1991-06

Fractals and chaos

Beaver, Philip Frederick

Monterey, California. Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

FRACTALS AND CHAOS

by

Philip Frederick Beaver

June 1991

Thesis Advisor: Maurice D. Weir
Co-Advisor: Ismor Fischer

Unclassified

ecurity Classification of this page

REPORT DOCUMENTATION PAGE

a Report Security Classification Unclassified		1b Restrictive Markings	
a Security Classification Authority		3 Distribution Availability of Report	
b Declassification/Downgrading Schedule		Approved for public release; distribution is unlimited.	
Performing Organization Report Number(s)		5 Monitoring Organization Report Number(s)	
a Name of Performing Organization Naval Postgraduate School	6b Office Symbol *(If Applicable)* MA	7a Name of Monitoring Organization Naval Postgraduate School	
c Address *(city, state, and ZIP code)* Monterey, CA 93943-5000		7b Address *(city, state, and ZIP code)* Monterey, CA 93943-5000	
a Name of Funding/Sponsoring Organization	8b Office Symbol *(If Applicable)*	9 Procurement Instrument Identification Number	
c Address *(city, state, and ZIP code)*		10 Source of Funding Numbers	
		Program Element Number / Project No / Task No / Work Unit Accession No	

1 Title *(Include Security Classification)* FRACTALS AND CHAOS

2 Personal Author(s) Beaver, Philip F.

3a Type of Report	13b Time Covered	14 Date of Report *(year, month,day)*	15 Page Count
Master's Thesis	From To	June 1991	179

6 Supplementary Notation The views expressed in this thesis are those of the author and do not reflect the official olicy or position of the Department of Defense or the U.S. Government.

7 Cosati Codes			18 Subject Terms *(continue on reverse if necessary and identify by block number)*
ield	Group	Subgroup	Fractal geometry and chaotic dynamical systems

9 Abstract *(continue on reverse if necessary and identify by block number*

The study of fractal geometry and chaotic dynamical systems has received considerable attention in the past ecade. Motivated by the interesting computer graphics produced by these fields, mathematicians have attempted o formalize the theoretical structure of the results, physicists have attempted to apply the theory to real world henomena, and laymen have enjoyed much of the popular literature and television programs the field has ostered. Unfortunately, the mathematics associated with these subjects has made them inaccessible to most ndergraduates, even if they have a strong background in mathematics. This thesis presents the basic ideas of ractal geometry and chaotic dynamical systems in a setting that can be understood by undergraduate students who ave had a course in advanced calculus. We hope it will allow them to gain an appreciation of the fields and notivate them to pursue further study.

0 Distribution/Availability of Abstract	21 Abstract Security Classification	
[X] unclassified/unlimited [ ] same as report [ ] DTIC users	Unclassified	
2a Name of Responsible Individual Maurice D. Weir	22b Telephone *(Include Area code)* (408) 646- 2608	22c Office Symbol MA/Wc

D FORM 1473, 84 MAR 83 APR edition may be used until exhausted

All other editions are obsolete

security classification of this page Unclassified

Approved for public release; distribution is unlimited.

Fractals and Chaos

by

Philip Frederick Beaver
Captain, United States Army
B.S., United States Military Academy, 1983

Submitted in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE IN APPLIED MATHEMATICS

from the

NAVAL POSTGRADUATE SCHOOL

ABSTRACT

The study of fractal geometry and chaotic dynamical systems has received considerable attention in the past decade. Motivated by the interesting computer graphics produced by these fields, mathematicians have attempted to formalize the theoretical structure of the results, physicists have attempted to apply the theory to real world phenomena, and laymen have enjoyed much of the popular literature and television programs that the field has fostered. Unfortunately, the mathematics associated with these subjects has made them inaccessible to most undergraduates, even if they have a strong background in mathematics. This thesis presents the basic ideas of fractal geometry and chaotic dynamical systems in a setting that can be understood by undergraduate students who have had a course in advanced calculus. We hope it will allow them to gain an appreciation of the fields and motivate them to pursue further study.

TABLE OF CONTENTS

I. INTRODUCTION 1

II. BASIC CONCEPTS 4

A. INTRODUCTION 4

B. METRIC SPACES 4

C. ITERATED FUNCTION SYSTEMS 6

D. CODE SPACE 13

E. THE CANTOR SET 15

III. FRACTALS 19

A. THE SETTING FOR FRACTALS 19

B. CONTRACTION MAPPINGS 22

C. AFFINE TRANSFORMATIONS OF THE PLANE 25

D. CONTRACTION MAPPINGS OF THE SPACE $\mathcal{H}(X)$ 29

E. CREATING FRACTALS THROUGH ITERATED FUNCTION SYSTEMS 33

1. The Cantor Set 33

2. Fractals in Two Dimensions 36

3. Condensation Sets 42

4. A Fractal Tree 43

F. APPLICATIONS OF FRACTALS TO COMPUTER GRAPHICS 46

G. THE ADDRESSES OF POINTS ON FRACTALS 50

H. FRACTAL DIMENSION 60

I. EXPERIMENTAL DETERMINATION OF FRACTAL DIMENSION 69

J. THE KOCH SNOWFLAKE 73

K. APPLICATIONS OF FRACTAL GEOMETRY 75

IV. CHAOS 77
A. INTRODUCTION 77
B. GRAPHICAL ANALYSIS OF FIXED POINTS OF MAPS 79
C. MAPS OF THE CIRCLE 82
D. CHAOTIC DYNAMICAL SYSTEMS 85
E. TOPOLOGICAL CONJUGACY 89
F. CHAOTIC DYNAMICS ON CODE SPACE 90
G. NEWTON'S METHOD FOR $X^2 = -1$ 92
H. THE QUADRATIC FAMILY OF MAPS 96
I. BIFURCATIONS 108
J. SARKOVSKII'S THEOREM 118
K. THE QUADRATIC FAMILY REVISITED 123
L. JULIA SETS 131
M. THE MANDELBROT SET 137
N. THE SMALE HORSESHOE 144
O. THE HENON MAP 149
P. THE LORENZ EQUATIONS 155

LIST OF REFERENCES 168

INITIAL DISTRIBUTION LIST 171

ACKNOWLEDGMENT

I would like to express my appreciation to a number of individuals who have played a role in the development of this thesis. I am particularly grateful to the following:

Professor Maurice D. Weir, who first suggested the subject of this thesis, and who developed the curriculum that provided the necessary background for me to write it. He continued to guide me throughout the preparation of the thesis, and spent countless hours encouraging, proofreading, and mentoring.

Professor Ismor Fischer, who led me through a course on chaotic dynamical systems, provided technical guidance and expertise during the preparation of this thesis, and who remained patient throughout the entire process.

Professor Aaron Schusteff, who started me on these subjects by teaching a course on fractal geometry and topology, and who remained an essential source of technical information during my further study of chaotic dynamical systems and throughout the preparation of this thesis.

Finally, Mr. David Beaver of The Automation Group, who introduced me to the idea of chaos by sending me James Gleick's book on the subject, and who volunteered numerous hours of word processing advice, without which I could not have prepared this document.

As usual, however, I retain sole responsibility for all the errors, mistakes, and other deviations from the truth that may be contained within this thesis.

Philip Beaver
Monterey, California
June 1991

I. INTRODUCTION

The subjects of fractals and chaos have attracted considerable attention in the last decade. This interest ranges from a "cult following" of laymen who are intrigued by the intricate computer graphics associated with the fields, to a rigorous mathematical treatment of the subjects by topologists and experts in dynamical systems, and to applications of these results to the real world by engineers and physicists. However, these subjects have been almost wholly inaccessible to undergraduates because of the level of mathematics required to study them. This thesis presents the subjects of fractals and chaos in a setting that can be understood by a typical undergraduate student with a solid background in mathematics through advanced calculus.

The subjects of fractals and chaos are not new. The German mathematician Georg Cantor (1845-1918) knew about fractals, and the French mathematician Jules Henri Poincare (1854-1912) knew about chaos in dynamical systems in the late nineteenth century. Additionally, the French mathematicians Pierre Fatou (1878-1929) and Gaston Julia (1893-1978) knew about Julia sets in the 1920s. However, it was not until the 1970s that high-speed computers allowed others to see what these men had discovered and to recognize the true potential of these fields. The growth of these fields has resulted in significant scientific advances in the past decade. While the discovery of quantum mechanics and relativity had a profound impact on very specialized areas of science, fractals and chaos have had a universal effect on the whole scientific community. Until recently, science had become so

specialized that, not only were physicists and mathematicians not communicating with each other, but molecular biologists were not communicating with population biologists either. The science of chaos has served to bring together the entire scientific community, including physicists, electrical engineers, mechanical engineers, biologists, economists, astronomers, meteorologists, medical scientists, and of course, mathematicians.

To explain briefly, fractals and chaos help to describe the universe. Natural objects, from crystals, plants and geological formations to weather patterns and galaxies, seem to have a fractal-like structure which eludes description by traditional geometric means. Additionally, dynamical (changing) systems in the real world, from turbulence in fluids to fluctuating economic trends and unpredictable weather patterns that previously defied description, are now being understood through the mathematics of chaos. While these fields are still in their infancy, the potential they have already demonstrated seems very promising in increasing our understanding of the physical world.

The mathematics required to understand this thesis includes basic courses in calculus and linear algebra. Additionally, due to some of the examples, a familiarity with numerical analysis is helpful, but not required. Most results (specifically in the fractals portion) are presented in the Euclidean plane, with a brief mention of more abstract spaces where applicable. Mathematical proofs that can be understood with this basic background are presented in their entirety. Results requiring more advanced mathematics are referenced. Where applicable, results are explained and interpreted in a basic setting. The

sections on Julia sets and the Mandelbrot set require a small amount of complex variable theory. Finally, only discrete dynamical systems are treated (with the exception of the section on the Lorenz equations), so a knowledge of differential equations is not required.

None of the material in this thesis constitutes original research. Instead, it is a synthesis of some of the best written works on the subjects of fractals and chaos, interpreted and presented in such a way as to be understood by a typical undergraduate with a basic mathematical background. Hopefully, it will serve to stimulate further interest and deeper study in these exciting new fields.

Specific references are listed in the sections to which they apply, and all of the figures and examples taken directly from references are so cited as they appear. However, we note that the two primary references for this thesis are Barnsley (1988) for the chapter on fractals, and Devaney (1989) for the chapter on chaos. Additionally, the references used for general information throughout the thesis are Ross (1980) for advanced calculus, Anton (1987) for linear algebra, Churchill (1990) and Boas (1987) for complex variables, and Giordano/Weir (1991) and Arnold (1987) for differential equations.

II. BASIC CONCEPTS

A. INTRODUCTION

There are several recurring mathematical concepts that appear in the study of both fractals and chaos. An understanding of these basic concepts simplifies this analysis. Instead of presenting them as they arise, we next discuss these concepts as background material. They are presented in a strict mathematical setting which is easily adapted to their further treatment within the settings of fractals and chaotic dynamical systems. This background material is common to most of the references cited (either as presented material or assumed knowledge), but the presentation here most closely follows those of Barnsley (1988) and Devaney (1989).

B. METRIC SPACES

We first introduce the concept of **metric spaces**. Although the notion of a metric space defined below may seem abstract and unfamiliar, the most common examples of the real line and the Euclidean plane are encountered in calculus. Many results from fractal geometry can be applied to general metric spaces, but for the purpose of simplifying them, they are presented in their most natural or familiar settings (such as the real line or the Euclidean plane).

A **space** is simply a collection of objects (a set) with some additional structure imposed on it. In a metric space, the additional structure is a **metric**, or **distance function**, which relates every two elements x and y in the set to a

unique nonnegative real number $d(x, y)$, called the **distance** between them. Any distance function (or metric) must satisfy the following three axioms for all x, y, z in the space:

1. **Symmetry:** $d(x, y) = d(y, x)$;
2. **Positive Definiteness:** $d(x, y) \geq 0$, and $d(x, y) = 0 \Leftrightarrow x = y$;
3. **Triangle Inequality:** $d(x, y) \leq d(x, z) + d(z, y)$.

A metric space is denoted by (X, d), where X is the set and d is the particular distance function. The metric space (R, d_E) is the real line R with $d(x, y) = |x - y|$. The space (R^2, d_E) is the Euclidean plane, where $X = R \times R$ and $d(x, y) = [(x_1 - y_1)^2 + (x_2 - y_2)^2]^{1/2}$. Whenever possible, we confine our examples to the real line or the Euclidean plane with these metrics.

We also need the concept of *closure*, a thorough treatment of which can be found in most advanced calculus texts. The **closure** of a set is the intersection of all closed sets containing that set. An equivalent definition, often presented as a theorem, is that if a sequence of points in a set S converges to a point x in the space, then x is in the closure of S. Thus, every set is a subset of its closure. As an example, consider the set Q of rational numbers. The number $\sqrt{2}$ is not in Q, but the sequence of rational numbers $1.4, 1.41, 1.414, 1.4142, \ldots$ converges to $\sqrt{2}$, so $\sqrt{2}$ is in the closure of Q (in fact, every real number is in the closure of Q).

A concept frequently associated with closure is that of *denseness*. A set S is **dense** in a metric space (X, d) if for every point $x \in X$ and all $\varepsilon > 0$, there exists a point $s \in S$ such that $d(s, x) < \varepsilon$. For example, the rational numbers form a dense subset of the real line. Another way to state this is that between

any two points on the real line, there exists a rational point. If a set S is dense in the space X, then the closure of S is X.

Much of our work on fractals and chaotic dynamical systems involves a topological property of some metric spaces called *completeness*. A **Cauchy sequence** is a sequence in which the terms get arbitrarily close to each other, i.e., for any $\varepsilon > 0$, there exists a number N such that $d(x_n, x_m) < \varepsilon$ for all $m, n > N$. A **complete metric space** is one where any Cauchy sequence of points in the space converges to a point in the space. An intuitive example is to again consider the space of rational numbers and the sequence 1, 1.4, 1.41, 1.414, 1.4142, . . . , which converges to the irrational $\sqrt{2}$. Clearly, terms in this sequence get arbitrarily close to each other; however, the limit value ($\sqrt{2}$) is not in the space of rational numbers, so the sequence does not converge to a point in the space. This is to say that there are "holes" in the space of rational numbers, hence it fails to be a complete space. The real line, on the other hand, is complete, since all Cauchy sequences of real numbers do converge to a real number. The spaces (R, d) and (R^2, d_E), are examples of complete metric spaces. A rigorous development of the concept of completeness and formal proofs of the completeness of the spaces just mentioned can be found in most advanced calculus texts.

C. ITERATED FUNCTION SYSTEMS

Fractals can be created, and chaotic dynamical systems understood, through an analysis of iterated function systems. A simple one-dimensional case of such a system is demonstrated by entering a starting value in a calculator and repeatedly activating a single function key (for example, the x^2

key). The result of executing this idea, with a starting value of 2 and the x^2 key, is $2, 2^2, 2^4, 2^8, 2^{16}, \ldots$. In this example, the system diverges to infinity and demonstrates a regular and predictable behavior. Another example of an iterated function system on a calculator is to use the $\sqrt{x}$ key and start with any positive number. Regardless of the initial value, this system will converge to 1. Unfortunately, not all iterated function systems are quite so simple. The study of chaotic dynamical systems attempts to explain irregular behavior in more complicated iterated function systems.

A **function**, or **map**, is a rule which assigns to each element in a specified domain a unique element in a codomain. The usual notation for this concept is $f: D \to R$, where f is the rule, D is the domain, and R is the codomain. A real-valued function of a real variable is specified by $f: \mathbf{R} \to \mathbf{R}$, and a function mapping the closed unit interval to itself is written $f: [0, 1] \to [0, 1]$. A function $f: \mathbf{R} \to \mathbf{R}$ of the form $f(x) = ax$ where a is a constant, is said to be **linear**; if it is of the form $f(x) = ax + b$ with a and b constants, it is called **affine**.

A function is **one-to-one** if each element in the range is the image of a unique element in the domain; that is $f(x) = f(y)$ implies $x = y$. A function is **onto** if every element in the codomain is the image of at least one element in the domain. A continuous function $f: \mathbf{R} \to \mathbf{R}$ is stated as $f \in C$, and a continuously differentiable function is stated as $f \in C^1$. A function that is n-times continuously differentiable is expressed by $f \in C^n$, and its nth derivative is written as $f^{(n)}(x)$. If a function is one-to-one, then its inverse f^{-1} exists according to the rule $f^{-1}(x) = y$ where $f(y) = x$. The domain of f^{-1} is the range of f. A function which is continuous, one-to-one, and onto for which f^{-1} is also continuous is called a **homeomorphism**. If f is a

homeomorphism and f and f^{-1} both are differentiable, then f is called a **diffeomorphism.**

As previously mentioned, an iterated function system results from repeatedly applying a function to an initial point x. The sequence x, f(x), f(f(x)), f(f(f(x))), . . . is also written as x, f(x), $f^2(x)$, $f^3(x)$, ... This sequence of iterates is called the **forward orbit of** x **under** f, and is denoted by $O^+(x)$. From our example above, the forward orbit of 2 under $f(x) = x^2$ is 2, 4, 16, 256, ... If f is a homeomorphism, f^{-1} exists, and the **backward orbit of** x **under** f is the sequence x, $f^{-1}(x)$, $f^{-2}(x)$, . . . , denoted by $O^-(x)$. Since $f(x) = x^2$ is not a homeomorphism, there is no backward orbit under f. However, if we let f(x) = 2x, then the backward orbit of 2 under f is 2, 1, 1/2, 1/4, ... A point x for which f(x) = x is called a **fixed point** of f; a point x for which $f^n(x) = x$ is called a **periodic point of period** n. The smallest positive integer n such that $f^n(x) = x$ is called the **prime period** of x. A point x is said to be **eventually periodic** if x is not periodic, but there exists an integer m such that $f^m(x)$ is a periodic point. For a function f where the derivative is defined on the entire domain, a point x is a **critical point** if f'(x) = 0, for example $f(x) = x^2$ at x = 0. It is a **degenerate critical point** if it is a critical point and f''(x) = 0, where the prime notation denotes differentiation of the function.

EXAMPLE. The following simple example illustrates a situation when a fixed point is guaranteed to exist for a given function. Let $f: [0, 1] \rightarrow [0, 1]$ be such that $f \in C$. Then f has at least one fixed point on [0, 1], as we now prove. If f(0) = 0 or f(1) = 1, then 0 or 1 is a fixed point and we are done. Otherwise,

$f(0) > 0$ and $f(1) < 1$, and define $g(x) = f(x) - x$. Since g is the difference of two continuous functions, g is continuous. Also $g(0) > 0$ and $g(1) < 0$. Hence, by the Intermediate Value Theorem from elementary calculus, there exists a point $a \in [0, 1]$ such that $g(a) = 0$. Therefore, $f(a) = a$, completing the proof (see Figure 2.1). This proof can be extended to any continuous function mapping a closed interval $[a, b]$ to itself.

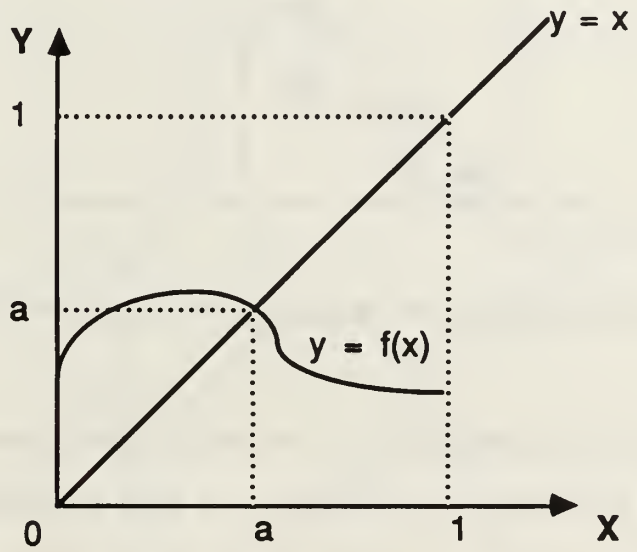

Figure 2.1 The fixed point of a graph f: $[0, 1] \rightarrow [0, 1]$.

We will frequently use a geometric or graphical analysis to investigate iterated function systems. For example, the orbit of two points, p and q, under $f(x) = x^2$ is shown graphically in Figure 2.2. Clearly, 1 and 0 are fixed points of $f(x) = x^2$. The orbits of points greater than 1 diverge to infinity, and

the orbits of points in the interval [0, 1) converge to 0. This example suggests the notion of fixed points (or periodic points) as being "attracting" or "repelling."

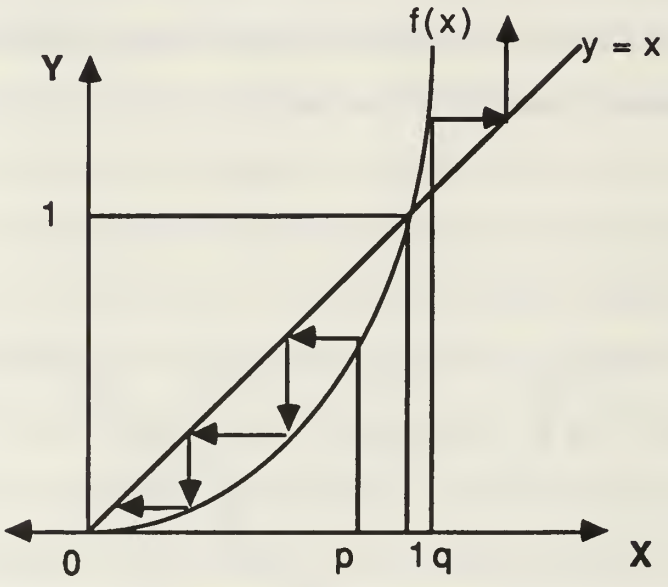

Figure 2.2 The orbits of two points p and q under $f(x) = x^2$.

A periodic point p of prime period n (to include fixed points with n = 1) is called **hyperbolic** if $|(f^n)'(p)| \neq 1$. If $|(f^n)'(p)| < 1$, then p is called an **attractor**, and if $|(f^n)'(p)| > 1$, then p is called a **repellor**. As in the above example with $f(x) = x^2$, the orbit of points near an attractor tend towards that point, while the orbits of points near a repellor tend away from that point. A precise definition of "near" will be given in the chapter on chaos. With $f(x) = x^2$, the derivative is $f'(x) = 2x$, so the derivative evaluated at the fixed points 0 and 1 gives $f'(0) = 0$ and $f'(1) = 2$. Therefore, 0 is an attracting fixed point and 1 is a repelling fixed point. Values in the interval [0, 1) tend

towards 0 and away from 1, while values in the interval $(1, \infty)$ all diverge. Hence, the point 0 "attracts" iterates and 1 "repels" them.

A graphical analysis can be very useful when analyzing fixed, periodic, or eventually periodic points. The graphs in Figures 2.3 through 2.6 show functions with fixed points, periodic points, eventually fixed points, and eventually periodic points. While geometric constructions cannot replace rigorous mathematical proofs, they are useful to demonstrate and provide insight into phenomena that occur in iterated function systems.

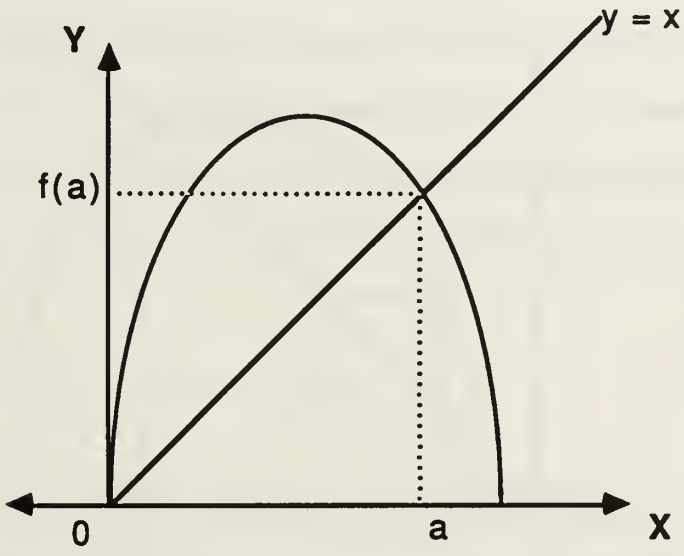

Figure 2.3 A fixed point a of f(x): f(a) = a.

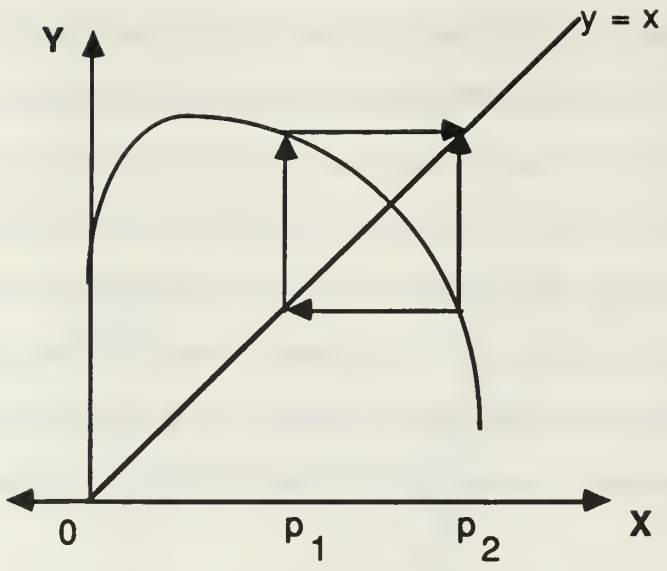

Figure 2.4 Periodic points of f(x): $f(p_1) = p_2$ and $f(p_2) = p_1$.

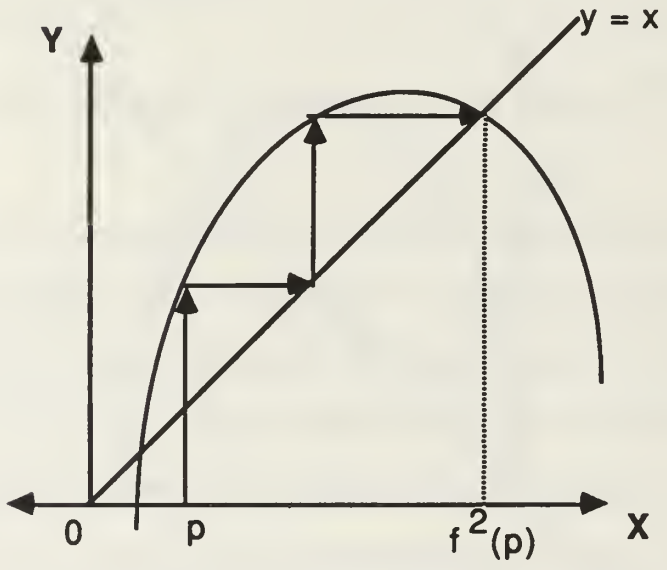

Figure 2.5 An eventually fixed point p of f(x): $f(f^2(p)) = f^2(p)$.

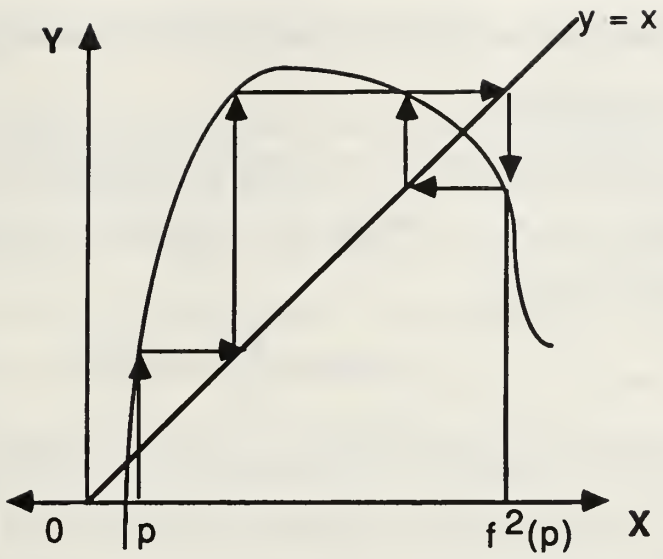

Figure 2.6 An eventually periodic point p of f(x): $f^2(p)$ is periodic.

D. CODE SPACE

A useful concept for identifying different points on fractals (frequently called *addressing points* on fractals) and analyzing chaotic dynamical systems is that of **code space.** We define Σ_2 as the set of all infinite sequences of binary digits $s_1s_2s_3\ldots$ where $s_i \in \{0, 1\}$. We next define a distance function for all $s, t \in \Sigma_2$ by

$$d(s, t) = \sum_{i=1}^{\infty} |s_i - t_i|/2^i.$$

Recalling the properties of a metric, it is easy to verify that this mapping does define a distance function on Σ_2 since $|s_i - t_i| \leq 1$. . Since the distance between any two points in Σ_2 is dominated by the convergent geometric series

$$\sum_{i=1}^{\infty} \frac{1}{2^i} = 1,$$

a unique and finite distance exists between any two points in the code space (that is, d(s, t) is well-defined).

The concept of code space can be extended to Σ_N, which consists of the infinite sequences $s_1, s_2, s_3, \ldots$, where $s_i \in \{0, 1, 2, 3, \ldots, N\text{-}1\}$. Here the distance function becomes

$$d(s, t) = \sum_{i=1}^{\infty} |s_i - t_i|/(N^i).$$

It is an easy exercise to show that the properties of a metric space are satisfied.

We will frequently perform an operation on code space known as the *shift map*. The **shift map** $\sigma: \Sigma_2 \to \Sigma_2$ is given by $\sigma(s_1s_2s_3\ldots) = (s_2s_3s_4\ldots)$. The shift map simply drops the first term in the sequence. Since $\sigma(0s_2s_3s_4\ldots) = \sigma(1s_2s_3s_4\ldots)$, the shift map fails to be one-to-one.

In an iterated function system of the shift map $\sigma, \sigma^1, \sigma^2, \sigma^3, \ldots$ on Σ_2, the only fixed points are $000\ldots$ and $111\ldots$; the eventually fixed points are of the form $s_1s_2\ldots s_n000\ldots$ and $s_1s_2\ldots s_n111\ldots$, and the periodic (or eventually periodic) points consist of sequences having a repeating (or eventually repeating) block of 0s and 1s. However, there are infinitely many points which are neither fixed nor periodic. These points are the sequences which have no repeating nor eventually repeating blocks of 0s and 1s.

We will use code space extensively and develop it in greater detail when discussing fractals and chaotic dynamical systems. For now, this introduction provides a brief exposure to the basic ideas of this important concept.

E. THE CANTOR SET

The Cantor set, which has traditionally served as an important pathological example in analysis, has secured a more distinct role in the studies of chaos and fractals. While it is actually one of the simplest examples of a one-dimensional fractal, we first present it in its traditional construction to ensure familiarization with the set before analyzing it in its new guise.

To construct the classical Cantor, or middle-thirds set, begin with the unit interval $[0, 1]$ and remove its middle third open interval $(1/3, 2/3)$. Of the two remaining line segments, $[0, 1/3]$ and $[2/3, 1]$, remove the open middle third of each of them. Of the four remaining line segments, again remove their open middle thirds, and continue this iterative procedure ad infinitum. The remaining set of points, which is a subset of the unit interval, is called the **Cantor middle-thirds (or ternary) set**. It is sometimes referred to as "Cantor dust" because of the scattered configuration of the remaining points. A picture of the construction of the classical Cantor set is shown in Figure 2.7.

The Cantor set clearly contains an infinite number of points because, at the very least, the endpoints of the line segments left after each iteration remain in the set. Nevertheless, the total amount of length removed from the unit interval is equal to one. To see this, consider the length removed during each iteration. In the first step, a segment of length $1/3$ is removed; in the second step, two segments each of length $1/9$ are removed; and in the third step, four segments each of length $1/27$ are removed. More generally, in the kth step, 2^{k-1} segments each of length $1/3^k$ are removed. Summing the total lengths removed yields

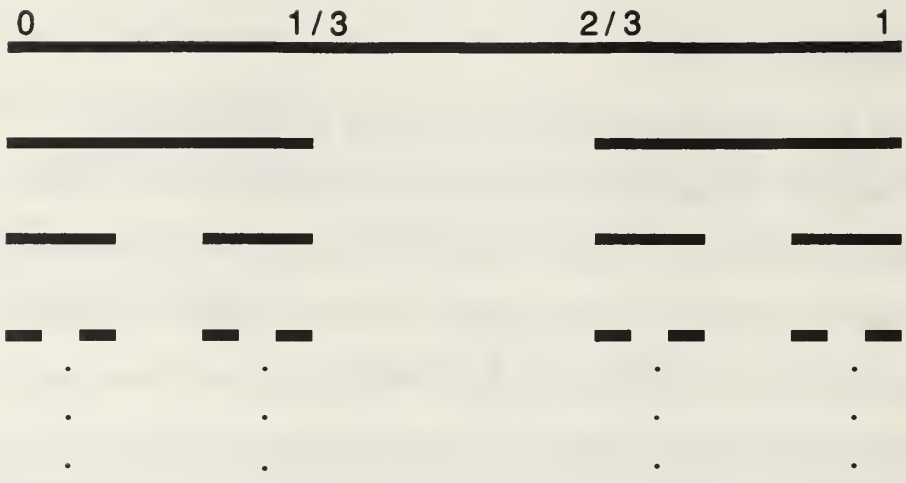

Figure 2.7 Constructing the classical Cantor set by removing the middle third through infinite iteration. The Cantor set consists of the points that remain.

$$L = \sum_{n=1}^{\infty} 2^{n-1}(1/3)^n = (1/3)\sum_{n=0}^{\infty} (2/3)^n = (1/3)[1/(1 - 2/3)] = 1.$$

So the total length removed is the entire length of the unit interval.

Another way of approaching the Cantor set is through ternary expansion. Consider expanding each point in the unit interval in its base three form $0.x_1x_2x_3\ldots$, where $x_i \in \{0, 1, 2\}$. The value of each point is then $x_1(1/3) + x_2(1/3^2) + x_3(1/3^3) + \ldots$. There is a minor technicality with this approach, because many points in the unit interval have dual ternary representations. For example, the point $1/3$ can be expressed both as $.1000\ldots$ and as $.0222\ldots$. In fact, the endpoints of every interval in our previous construction of the Cantor set, except for 0 and 1, have such a dual representation. The way we remedy this is to always use the representation containing the repeating 2. Now, all of the points between $1/3$ and $2/3$ have $x_1 = 1$, and these are the

points removed in the first step of our former construction. Similarly, the points removed in the second step have the form $x_1 = 0$ or 2, and $x_2 = 1$. Continuing with this analysis, we see that the points remaining in the Cantor set are precisely those that have no 1s appearing in their ternary expansion.

Thus the Cantor set consists of all points in the unit interval having only 0s or 2s in their ternary expansion. Moreover, every point which has only 0s or 2s in its ternary expansion is a member of the Cantor set. This definition proves to be very useful when we analyze chaotic dynamical systems. Note that the Cantor set contains no interval subsets because between any two points, there is a point with a 1 in its ternary expansion. To see this, assume that the Cantor set does contain an interval subset. We can let the endpoints of this interval be $x_1x_2\ldots x_n0x_{n+2}\ldots$ and $x_1x_2\ldots x_n2x_{n+2}\ldots$, where the x_i agree for the first n digits, are different for the n+1st digit (which means one must be a 0 and the other must be a 2), and after which the digits can be arbitrary. Since every point between these two must be in the Cantor set, points of the form $x_1x_2\ldots x_n1\ldots$ must be in the set, contradicting the fact that the Cantor set contains no points with 1s in their ternary expansion. This proves that the Cantor set contains no intervals, thus its only "connected" subsets are single points. (We discuss the concept of *connectedness* in detail in the chapter on Julia sets). This illustrates another property of the Cantor set: it is **totally disconnected**.

Now a great paradox results. We have removed the entire length of the interval. The only points remaining are the points of the form $x = .x_1x_2x_3\ldots$ where $x_i \in \{0, 2\}$. Now we form the function $f(x) = y$ where $y_i = x_i/2$. The set of y values is the set of *all strings* of 0s and 1s. The cardinality of this set

is the same as the cardinality of [0, 1] since there is a one-to-one correspondence with the binary expansion of the points of the unit interval! Thus, the cardinality of a set does not tell the whole story of "how many" points are actually there from the point of view of length (in the one-dimensional case).

There are other versions of Cantor sets resulting from similar constructions. For instance, remove instead the middle fourth, fifth, or some other fixed (1/N) length at each step. We can also remove from the unit interval an open middle segment of length $\alpha/3^k$ at the kth stage where $0 < \alpha < 1$ is fixed. Then we can show the total length removed is α. These are frequently called "fat" Cantor sets, because the total length removed from the unit interval is now less than one. These sets can also be described in terms of a base N expansion, from which it can be shown that a fat Cantor set contains no intervals.

III. FRACTALS

A. THE SETTING FOR FRACTALS

Fractal geometry involves the study of certain subsets of metric spaces. It can be viewed as an extension of Euclidean geometry and is frequently used to describe objects that occur in nature, such as crystals, plants, clouds, and geological formations (see, for example, Cherbit (1991)). Fractals have been applied to computer graphics to store information efficiently, examples of which can be found in Barnsley (1988). Moreover, the use of fractals to study real-world phenomena has provided a new way of analyzing the world. Finally, as we shall see in Chapter IV, a primary use of fractals is to classify and analyze chaotic dynamical systems.

The primary references for this chapter are Barnsley (1988) and Falconer (1990). While the presentation most closely follows that of Barnsley, a more rigorous mathematical development of most of the results can be found in Falconer. Additionally, the article by Harrison (1989) covers much of the same material, while many of the examples presented here are from Cherbit (1991).

Much of the current literature differs in the precise definition of a fractal, so our approach is to develop the setting for the space in which fractals exist. Then we provide many examples of fractals in that setting. This approach provides a good initial understanding of fractals without the expense involved in achieving a thorough understanding of their elusive definition.

A useful, although incomplete, definition of a fractal is that "a fractal is a fixed point of a certain kind of transformation on the space $(\mathcal{H}(X), h(d))$." Thus, we must first define the set $\mathcal{H}(X)$ and the distance function $h(d)$. $(\mathcal{H}(X), h(d))$ is a metric space obtained from a complete metric space (X, d), where the points in $\mathcal{H}(X)$ are closed, bounded, nonempty subsets of X, and where $h(d)$ is a distance function based on the metric d, (which we define shortly). The points in $\mathcal{H}(X)$ are called the "compact" subsets of X. To avoid the necessity of employing concepts from advanced calculus, we consider only Euclidean spaces where the compact sets are the closed and bounded subsets.

Consider the Euclidean plane $(\mathbf{R}^2, d_E)$. We denote by $\mathcal{H}(\mathbf{R}^2)$ the collection of all closed and bounded subsets of $\mathbf{R}^2$, excluding the empty set. Hence, the closed unit square $[0, 1]X[0, 1]$ belongs to $\mathcal{H}(\mathbf{R}^2)$, as does the origin; however, the interval $(0, 1)X\{0\}$ is not an element in $\mathcal{H}(\mathbf{R}^2)$ since it is not a closed subset of $\mathbf{R}^2$. Clearly, the union of any two elements in $\mathcal{H}(\mathbf{R}^2)$ again belongs to $\mathcal{H}(\mathbf{R}^2)$. The intersection of two elements in $\mathcal{H}(\mathbf{R}^2)$ is not necessarily an element of $\mathcal{H}(\mathbf{R}^2)$ as the intersection may be empty.

In order to create a metric space out of the set $\mathcal{H}(\mathbf{R}^2)$, a distance function (metric) is required that relates any two elements of $\mathcal{H}(\mathbf{R}^2)$ to a nonnegative real number. To this end the concept of dilation is helpful. Given any closed subset A of $\mathbf{R}^2$ and $\varepsilon > 0$, the **ε-dilation** of A is defined to be the set of all points x in $\mathbf{R}^2$ such that the smallest Euclidean distance between x and any point in the set A is less than or equal to ε; i.e., the ε-dilation of A is the set $\{x: d_E(x, A) \leq \varepsilon\}$. For example, the ε-dilation of the origin is simply the closed disc with radius ε. The ε-dilation of the unit square $[0, 1]X[0, 1]$ with $\varepsilon = 1/2$ is shown in Figure 3.1. The points of the square also belong to its ε-dilation.

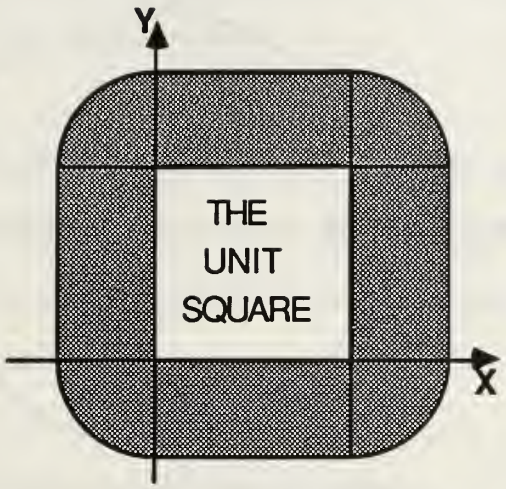

Figure 3.1 The ε-dilation of the unit square with $\varepsilon = 1/2$.

We now define the distance from an element A in $\mathfrak{H}(R^2)$ to an element B in $\mathfrak{H}(R^2)$ as the number $d_h(A, B)$ representing the smallest ε such that every point in B is covered by the ε-dilation of the element A. In other words, $d_h(A, B) = \min(\varepsilon: y \in \{\text{the } \varepsilon\text{-dilation of A}\}\ \forall\ y \in B)$. Clearly, d_h exhibits the second and third properties of a metric from the original definition. However, as shown in Figure 3.2, it is not symmetric since $d_h(A, B) \neq d_h(B, A)$ in general. To remedy this difficulty, we define $h(A, B) = \max\{d_h(A, B), d_h(B, A)\}$. The number $h(A, B)$ does satisfy the properties of a metric for points in the set $\mathfrak{H}(R^2)$, and is referred to as the **Hausdorff distance**. Now the set $\mathfrak{H}(R^2)$ together with the metric $h(A, B)$ is a metric space. It is denoted by $(\mathfrak{H}(R^2), h(d))$, or for purposes of simplicity just $\mathfrak{H}$, and we refer to it as the **Hausdorff-metric space**. We remark that if the metric space (X, d) is complete, then the associated Hausdorff-metric space $(\mathfrak{H}(X), h(d))$ is also complete. In particular, $\mathfrak{H}$ is complete since $X = R^2$ is complete.

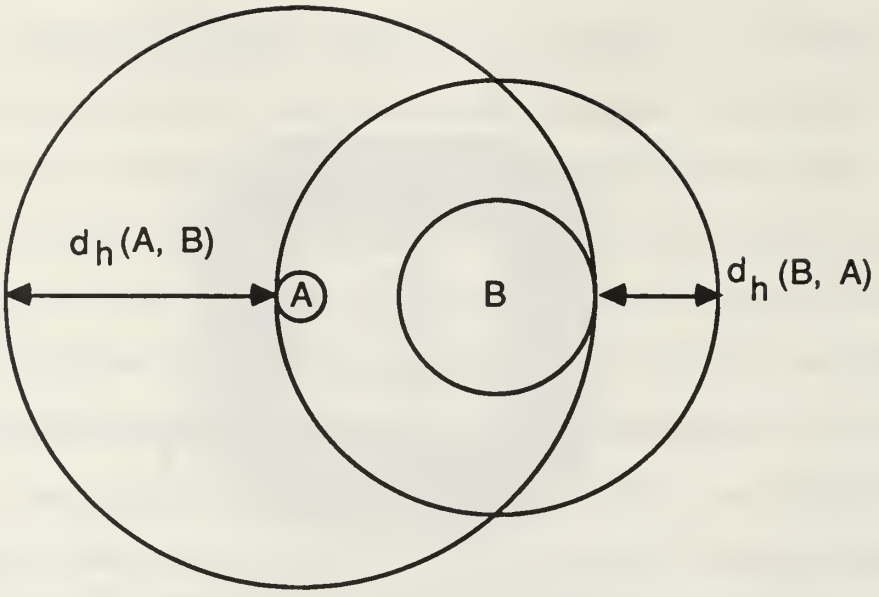

Figure 3.2 The distances $d_h(A,B)$ and $d_h(B,A)$ are not equal in general.

B. CONTRACTION MAPPINGS

In order to generate fractals geometrically we need to be familiar with the concept of contraction mappings. The results of this section are presented in the setting of a general metric space. However, we concentrate our efforts on three main spaces: the real line with the normal distance function, the Euclidean plane, and the Hausdorff-metric space $(\mathcal{H}(\mathbf{R}^2), h(d_E))$.

A mapping $f: X \to X$ is said to be a **contraction mapping** if there exists a constant $0 \le s < 1$ such that $d(f(x), f(y)) \le s(d(x, y))$ for all $x, y \in X$. The number s is called the **contractivity factor** for f. We prove shortly that a contraction mapping is always continuous. Under a contraction, any two points in the space that begin a distance D apart will be moved to within a distance sD of each other. A key result concerning contraction mappings,

and one which is critical when constructing fractals as subsets of the Hausdorff-metric space $\mathfrak{H}$, is discussed next.

THE CONTRACTION MAPPING THEOREM.

If $f: X \to X$ is a contraction defined on a complete metric space (X, d), then f has a unique fixed point $x_f \in X$. Furthermore, for any point $x \in X$, the sequence $f^n(x)$ converges to x_f.

To prove this we first need the following results.

THEOREM. If f is a contraction, then it is continuous.

PROOF. If we let s be the contractivity factor of f, and if $\varepsilon > 0$ is given, then choosing $\delta = \varepsilon/s$ yields $d(x, y) < \delta \Rightarrow d(f(x), f(y)) \leq s(d(x, y)) < s(\varepsilon/s) = \varepsilon$. This shows that f is continuous.

LEMMA. If f is a contraction with contractivity factor s, then for a fixed x and $m < n$, $d(f^n(x), f^m(x)) \leq s^m d(x, f^{n-m}(x))$.

PROOF. The proof follows immediately from the contractivity factor s, and the principle of mathematical induction.

PROOF OF THE CONTRACTION MAPPING THEOREM.

Let x_0 be an arbitrary point in the complete metric space (X, d), and let $f: X \to X$ be a contraction mapping such that $x_1 = f(x_0)$, $x_2 = f(x_1) = f^2(x_0)$, and in general, $x_n = f(x_{n-1}) = f^n(x_0)$. For positive integers m and n such that $m < n$, we have

$$d(x_m, x_n) = d(f^m(x_0), f^n(x_0)) = d(f^m(x_0), f^m(f^{n-m}(x_0)))$$

$$\leq s^m d(x_0, f^{n-m}(x_0)) = s^m d(x_0, x_{n-m})$$

$$\leq s^m[d(x_0, x_1) + d(x_1, x_2) + \ldots + d(x_{n-m-1}, x_{n-m})]$$

(by the triangle inequality of d)

$$\leq s^m d(x_0, x_1)[1 + s + s^2 + \ldots + s^{n-m-1}]$$

(by applying the lemma to each term)

$$< [s^m/(1 - s)]d(x_0, x_1)$$ (by the geometric series).

Since $s < 1$, $s^m \to 0$, hence $d(x_m, x_n) \to 0$. Since $\mathbf{X}$ is a complete metric space, $x_n \to x_f$ for some $x_f \in \mathbf{X}$, so the sequence $\{f^n(x_0)\}$ converges to the point x_f for any x_0 in the space. To show that x_f is a fixed point: since f is a continuous function, $f(x_f) = f(\lim x_n) = \lim f(x_n) = \lim x_{n+1} = x_f$. Therefore, x_f is a fixed point of f.

To show that x_f is the only fixed point of f, assume that y is also a fixed point. Then $d(x_f, y) = d(f(x_f), f(y)) \leq s\,d(x_f, y)$. Since $s < 1$ and $d \geq 0$, this implies that $d(x_f, y) = 0$, so $x_f = y$, completing the proof.

The importance of the contraction mapping theorem will become apparent when fractals are created through iterated function systems. When we find a point in the space that is a fixed point under a contraction mapping, we then know it is unique (hence we will refer to "the" fixed point of the mapping). Furthermore, we know that *any* initial point converges to the fixed point under iteration of f.

C. AFFINE TRANSFORMATIONS OF THE PLANE

Affine mappings were defined earlier as functions of the form $f(x) = ax + b$. In R^2, we change our notation slightly and write an affine transformation as $w(x) = Ax + t$, where x is a point in R^2 (a column two-vector), A is a 2X2 matrix, and t is a (fixed) vector in R^2. We frequently use the form

$$w(x) = w\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} = Ax + t.$$

The mapping $w(x) = Ax$ is called a **linear transformation**, and it takes any parallelogram with one vertex at the origin into another parallelogram with a vertex at the origin, provided that the determinant of A is not zero (that is, A is nonsingular). Notice that the origin remains fixed under any linear transformation. A result from linear algebra is that the matrix A can be written in the form

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} r_1\cos\theta_1 & -r_2\sin\theta_2 \\ r_1\sin\theta_1 & r_2\cos\theta_2 \end{bmatrix} \qquad (1)$$

where the point (a, c) has polar coordinates (r_1, θ_1) and the point (b, d) has polar coordinates $(r_2, \theta_2 + \pi/2)$. The transformation $w(x) = Ax$ "deforms" the space R^2 relative to the origin. By deform, we refer to the result from linear algebra that any linear transformation can be expressed as a succession of shears, compressions, expansions, and rotations of the space. The result of adding the vector t to a point in R^2 shifts that point by the magnitude and direction of t, so an affine transformation of the plane can be thought of as a

deformation (shearing, shrinking/stretching and rotating) followed by a translation. Figure 3.3 shows the result of the affine transformation

$$w\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

on the triangle T with vertices A = (0, 1), B = (1, 1), and C = (1, 0), where A', B', and C' are the images of A, B, and C under w, respectively.

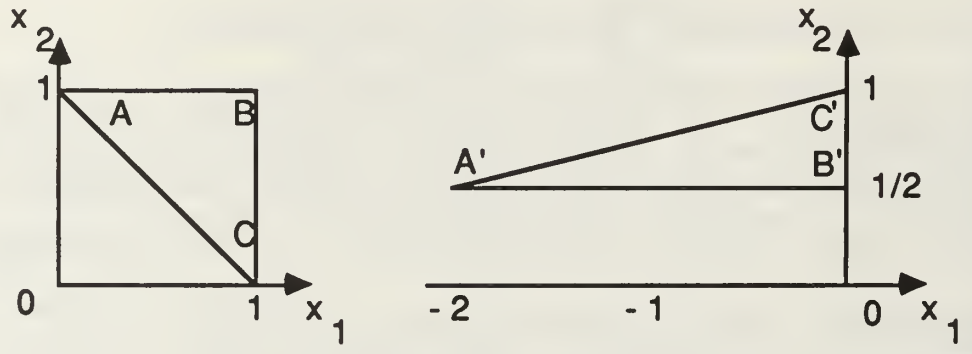

Figure 3.3 An affine transformation of a triangle.

We now consider several special affine transformations. From the representation in polar coordinates Eq. (1) above, if $r_1 = r_2 = r$, and $\theta_1 = \theta_2 = \theta$, we say the transformation is a **similitude**, and write it in the form

$$w\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} r\cos\theta & -r\sin\theta \\ r\sin\theta & r\cos\theta \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix},$$

or,

$$w\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} r\cos\theta & r\sin\theta \\ r\sin\theta & -r\cos\theta \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}.$$

A similitude scales any image in R^2 by the same factor r in each component direction, rotates it by the angle θ, and then translates it by the magnitude and direction of the vector t. Recall from linear algebra that the transformation

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is a **rotation**, and the transformation

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

is a **reflection** across the x-axis in R^2. Both of these transformations behave as one would expect, based on their names. Figure 3.4 shows the transformation

$$w\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos(\pi/2) & -\sin(\pi/2) \\ \sin(\pi/2) & \cos(\pi/2) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{2}$$

of the unit square which is a rotation with $\theta = \pi/2$. Figure 3.5 shows the transformation

$$w\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{3}$$

of the unit square which is a reflection about the x axis. Being familiar with both reflections and rotations is very useful when creating fractals.

The action of $w(x) = Ax + t$ can be determined by evaluating its effect on any point (x_1, x_2) in the plane. However, it is sometimes more useful to determine which affine transformation has caused a particular (observed) change to a given set of points. The beauty of affine transformations is that they are uniquely determined by their action on any three non-collinear

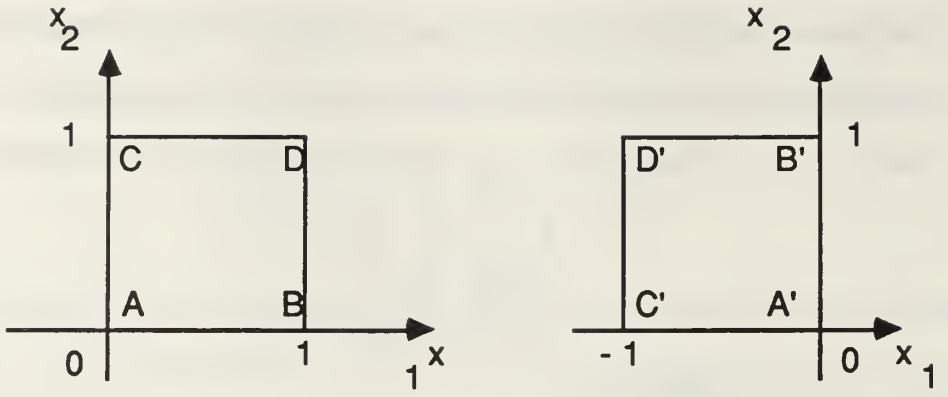

Figure 3.4 A rotation of the unit square by the transformation in Eq. (2).

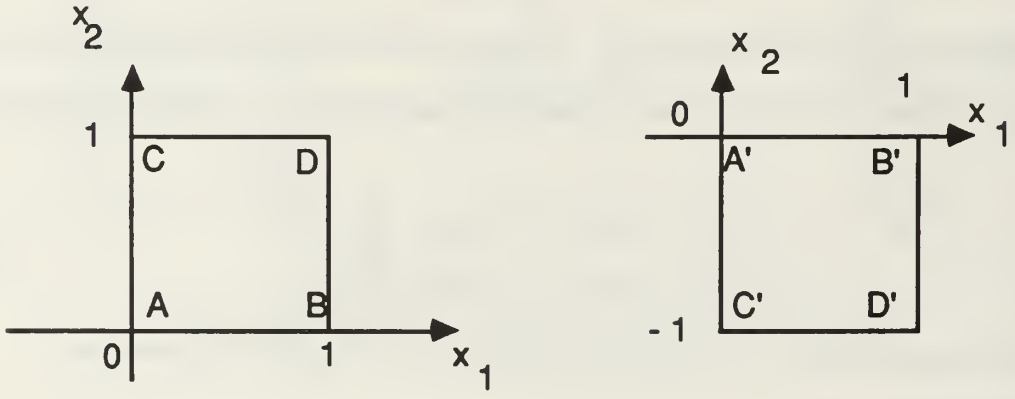

Figure 3.5 A reflection of the unit square by the transformation in Eq. (3).

points in the plane. To see this, start with any three points (x_1, x_2), (y_1, y_2), and (z_1, z_2) and note their movement under an (unknown) affine transformation to the new points (x_1', x_2'), (y_1', y_2'), and (z_1', z_2'). The coefficients a, b, c, d, e, and f of the transformation can be determined from the system of linear equations

$$x_1a + x_2b + e = x_1',$$

$$y_1a + y_2b + e = y_1',$$

$$z_1a + z_2b + e = z_1',$$

$$x_1c + x_2d + f = x_2',$$

$$y_1c + y_2d + f = y_2',$$

$$z_1c + z_2d + f = z_2'.$$

Because the points are non-collinear, the matrix of the coefficients is nonsingular and the system has a unique solution. An example follows.

EXAMPLE. Using the example in Figure 3.3, the transformation on the triangle transforming the points (1, 0), (1, 1), and (0, 1) to (-2, 1/2), (0, 1/2), and (0, 1) respectively yields the following system of equations:

$$(0)a + (1)b + e = -2,$$

$$(1)a + (1)b + e = 0,$$

$$(1)a + (0)b + e = 0,$$

$$(0)c + (1)d + f = 1/2,$$

$$(1)c + (1)d + f = 1/2,$$

$$(1)c + (0)d + f = 1.$$

Solving for the coefficients a through f yields $a = 2$, $b = 0$, $c = 0$, $d = -1/2$, $e = -2$, and $f = 1$, hence, the affine transformation is given by

$$w\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

D. CONTRACTION MAPPINGS OF THE SPACE $\mathfrak{H}(X)$

Given n continuous maps of a metric space $w_1, w_2, \ldots, w_n: X \to X$, we construct a map W of the associated space $\mathfrak{H}$ by considering

$$W(x) = w_1(x) \cup w_2(x) \cup \ldots \cup w_n(x),$$

where x is any point in the space $\mathcal{H}$, i.e., x is a closed, bounded, nonempty subset of the original metric space. For $x \in \mathcal{H}$, we define $w_i(x) = \{w_i(y): y \in x \subset X\}$. We state here that the mapping W maps $\mathcal{H}$ to itself. This fact is based on two important results from advanced calculus, which we state but do not prove. The first result is that continuous images of compact sets are compact, so that each w_i maps $\mathcal{H}$ to itself. The second is that the finite union of compact sets is compact. Thus, W maps $\mathcal{H}$ to itself, as claimed. For a simple illustration of this concept, let $n = 2$ and $X = \mathbf{R}$, and consider the maps on $\mathbf{R}$ defined by $w_1(x) = x$ and $w_2(x) = 2x$. If we start with the singleton compact set $x = \{1\}$, we have $W(1) = w_1(\{1\}) \cup w_2(\{1\}) = \{1\} \cup \{2\} = \{1, 2\}$.

We next iterate the map W. The set of points $W^n(x)$ grows quite rapidly as n increases. To see this, consider the same example used above with $n = 2$. Then $W^2(x) = W(W(x)) = W(w_1(x) \cup w_2(x)) = w_1(w_1(x) \cup w_2(x)) \cup w_2(w_1(x) \cup w_2(x)) = w_1(w_1(x)) \cup w_1(w_2(x)) \cup w_2(w_1(x)) \cup w_2(w_2(x))$. Using w_1 and w_2 as above, and again starting with $x = \{1\}$, we have $W^2(\{1\}) = W(\{1, 2\}) = \{1, 2\} \cup \{2, 4\} = \{1, 2, 4\}$. Similarly, $W^3(\{1\}) = \{1, 2, 4\} \cup \{2, 4, 8\} = \{1, 2, 4, 8\}$. Note that $w_1(w_2(x)) = w_2(w_1(x))$ because w_1 and w_2 are linear transformations on $\mathbf{R}$, which commute. For notational convenience, we shall not distinguish between the notation $x = 1$ and $x = \{1\}$ for the space $\mathcal{H}(\mathbf{R})$ includes sets of singleton points.

We use the abbreviated notation $w_i(w_j(x)) = w_{ij}(x)$ etc. Then, $W^1 = w_1 \cup w_2$; $W^2 = w_{11} \cup w_{12} \cup w_{21} \cup w_{22}$, and $W^3 = w_{111} \cup w_{112} \cup w_{121} \cup w_{122} \cup w_{211} \cup w_{212} \cup w_{221} \cup w_{222}$. These subscripts may look familiar. We are building, in this way, an iterated function system in one-to-one correspondence with the code space Σ_2. To make this clear, we let each 1 in

our subscripts correspond to the 0 in code space, and let each 2 in our subscripts correspond to the 1 in code space. Then, the iterate $w_{ijk...}$ is in direct correspondence with the point i-1, j-1, k-1, . . . in Σ_2. We could have emphasized this more had we named our contractions w_0 and w_1. However, we save the designation w_0 for a specific use later on. Similarly, if W is the union of n mappings, then the subscripts are analogous to the points in Σ_n after infinite iteration. This concept proves quite useful when we identify points on fractals through an addressing scheme.

Notice that the infinite iteration $W^{\infty}(1)$ in our simple example produces the set {1, 2, 4, 8, 16, . . .}. While the dynamics of this particular iterated function system are not very exciting, when each w_i is a contraction with associated contractivity factor s_i the results do become much more interesting, as we discuss next.

To discuss contraction mappings on the space $\mathfrak{H}$, we need the following two theorems.

THEOREM. If a mapping on a metric space $w: X \to X$ is a contraction with contractivity factor s, then $w: \mathfrak{H}(X) \to \mathfrak{H}(X)$ is also a contraction with contractivity factor s.

PROOF. Let $B, C \in \mathfrak{H}$. Then

$$\begin{aligned} d_h(w(B), w(C)) &= \min\{\varepsilon:\ y \in \{\text{the } \varepsilon\text{-dilation of } w(B)\}\ \forall\ y \in w(C)\} \\ &\le \min\{s\varepsilon:\ y \in \{\text{the } \varepsilon\text{-dilation of } B\}\ \forall\ y \in C\} \\ &= s \min\{\varepsilon:\ y \in \{\text{the } \varepsilon\text{-dilation of } B\}\ \forall\ y \in C\} \\ &= s\, d_h(B, C). \end{aligned}$$

Similarly, $d_h(w(C), w(B)) \le s\, d_h(C, B)$.

Therefore, $h(w(B), w(C)) \leq s\, h(B, C)$, completing the proof.

THEOREM. Let $w_1, w_2, \ldots, w_n$ be contractions on the space $\mathcal{H}$ with contractivity factors $s_1, s_2, \ldots, s_n$. Define $W: \mathcal{H} \to \mathcal{H}$ by

$$W(B) = w_1(B) \cup w_2(B) \cup \ldots \cup w_n(B)$$

for each $B \in \mathcal{H}$. Then W is a contraction mapping on $\mathcal{H}$ with contractivity factor $s = \max\{s_i: i = 1, 2, \ldots, \}$.

PROOF. We prove this for the case $n = 2$. The rest follows by induction. Let $B, C \in H$. Then

$$\begin{aligned} H(W(B), W(C)) &= h(w_1(B) \cup w_2(B), w_1(C) \cup w_2(C)) \\ &\leq \max\{h(w_1(B), w_1(C)), h(w_2(B), w_2(C))\} \\ &\leq \max\{s_1 h(B, C), s_2 h(B, C)\} = s h(B, C). \end{aligned}$$

A simple exercise shows why the first inequality holds. This completes the proof.

Figure 3.6 shows two contractions on the unit square in $\mathbf{R}^2$, w_1 contracts by $1/2$ in the x_1 direction, and w_2 contracts by $1/4$ in the x_2 direction. Their union has a contractivity factor of $1/2$ by the previous theorem.

Since W is a contraction on the complete metric space $\mathcal{H}(X)$ we know that it has a unique fixed point in $\mathcal{H}(X)$, which is a closed, bounded, non-empty subset of the space X. By the Contraction Mapping Theorem, all points in $\mathcal{H}$ tend to this fixed point under infinite iteration of W. This unique fixed point is called the **attractor** of W, and frequently exhibits very interesting properties.

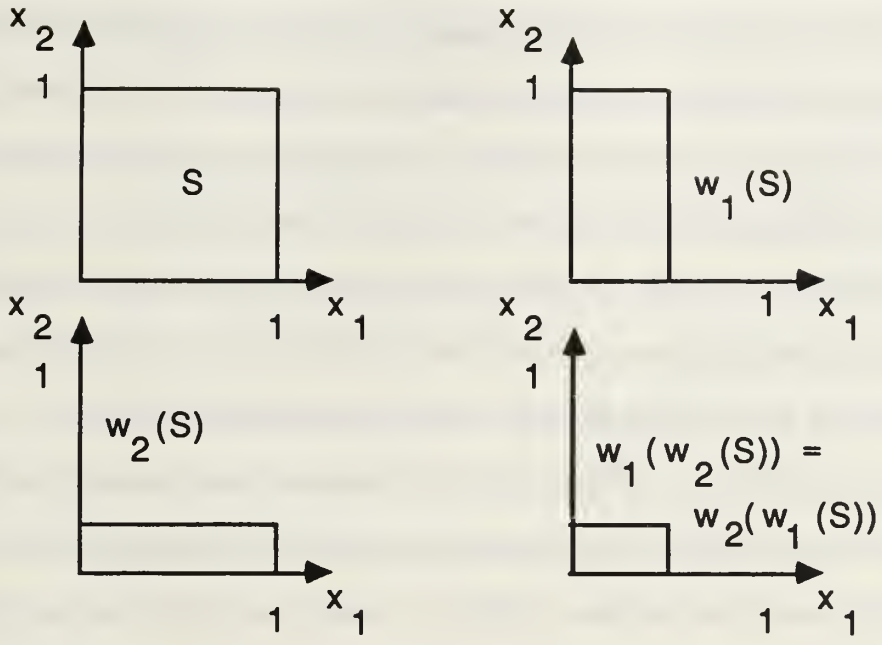

Figure 3.6 The union of two contractions of the unit square.

E. CREATING FRACTALS THROUGH ITERATED FUNCTION SYSTEMS

1. The Cantor Set

We are now prepared to construct fractals using affine transformations and the Contraction Mapping Theorem. We start with the simplest setting, namely the unit interval, and construct a fractal with which we are already familiar.

Consider the union of two affine transformations of the real line R, given by $W(x) = w_1(x) \cup w_2(x)$, where $w_1(x) = x/3$, and $w_2(x) = x/3 + 2/3$. If we first restrict our attention to the unit interval [0, 1], then under the first iteration of w_1 the unit interval is shrunk by 1/3 towards the origin such that $[0, 1] \rightarrow [0, 1/3]$. Under the first iteration of w_2 the unit interval is shrunk by 1/3 and then translated to the right by 2/3 so that $[0, 1] \rightarrow [2/3, 1]$. Note

that the effect of W on [0, 1] is the removal of the open middle third of the unit interval, the first step in constructing the Cantor "middle thirds" set.

Under the second iteration W^2, the mappings w_1 and w_2 now act on the subsets [0, 1/3] and [2/3, 1] of the unit interval, so that $W^2[0, 1] = w_{11}[0, 1] \cup w_{12}[0, 1] \cup w_{21}[0, 1] \cup w_{22}[0, 1]$. This can be written as $w_1[0, 1/3] \cup w_1[2/3, 1] \cup w_2[0, 1/3] \cup w_2[2/3, 1]$. Since $w_1[0, 1/3] = [0, 1/9]$, $w_1[2/3, 1] = [2/9, 1/3]$, $w_2[0, 1/3] = [2/3, 7/9]$, and $w_2[2/3, 1] = [8/9, 1]$, we obtain $W^2[0, 1] = [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1]$. This procedure has the same effect as removing the open middle third of each interval [0, 1/3] and [2/3, 1].

Using a similar analysis, $W^3[0, 1]$ again removes the open middle third of each of the four intervals produced from W^2, and continuing in this manner results in the construction of the Cantor middle-thirds set (see Figure 3.7).

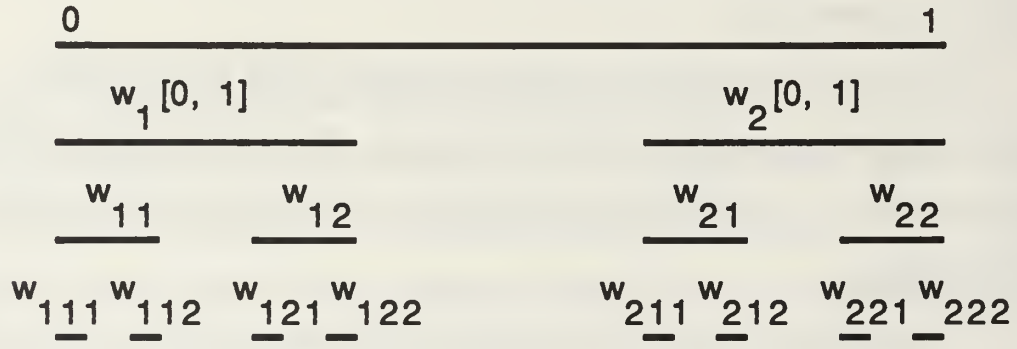

Figure 3.7 Creating the Cantor set through an iterated function system (IFS).

The above construction is perhaps the simplest example of a fractal. Further examination reveals that this set has some interesting properties. First, consider the set of points that are fixed under the map W. Note that the

fixed point of w_1 is 0, since $(1/3)(0) = 0$, and that the fixed point of w_2 is 1, since $(1/3)(1) + 2/3 = 1$. Since each of these mappings is a contraction of the real line, each has a unique fixed point by the contraction mapping theorem. However, under the union of these maps $W = w_1 \cup w_2$, the endpoints of the intervals we constructed, $1/3, 2/3, 1/9, 2/9, 7/9, 8/9, 1/27, \ldots$, also remain in the set after each iteration of W. While we restricted our attention to the unit interval, we note that each contraction w_i is defined on the real line and their union is a contraction on the space $\mathcal{H}(\mathbf{R})$. Hence, if we can find a closed, bounded subset of the real line whose image under W is itself, we know it is the unique fixed point of the mapping W. The classical Cantor set is such a point, because taking the contraction W of the Cantor set again yields the Cantor set. To see this, recall that in our construction of the Cantor set each subinterval contains a smaller copy of the whole set because of the infinite iteration we used. Likewise, since W maps the unit interval to $[0, 1/3]$ and $[2/3, 1]$, each subinterval also contains a smaller copy of the whole interval. Since the Cantor set is the fixed point of this iterated function system (IFS), we refer to it as the attractor of W.

This illustrates a property of fractals known as **self-similarity**. If we examine the interval $[0, 1/3]$ in the construction of the Cantor set, we see that it is a scaled copy of the original unit interval. Likewise, if we examine any subinterval of the unit interval, we observe the same result. This is a unique property of the infinite iteration scheme used to construct the Cantor set. Similarly it is the result of iterating W an infinite number of times. This self-similarity characteristic of fractals is very useful in applications to computer graphics.

Because it was known ahead of time that the attractor of W would be contained entirely within the unit interval, to simplify the discussion we based our previous analysis on mappings of the unit interval $W: [0, 1] \rightarrow [0, 1]$. Now consider the result of mapping a different compact (i.e., closed and bounded) subset of the real line. From the Contraction Mapping Theorem, no matter which nonempty compact subset of the real line we begin with, infinite iteration of W always produces the Cantor set. Although we proved this principle formally, it certainly is not intuitively obvious. In the next section we illustrate this idea with an example in $\mathbf{R}^2$, where the geometry can be better appreciated.

The question now arises, what happens to the entire real line under iteration by W? It is clear that $W(-\infty, \infty) = (-\infty, \infty)$, which appears to be a fixed point of W. However, the set $(-\infty, \infty)$ is not in the space $\mathcal{H}(\mathbf{R})$ since points in the space $\mathcal{H}(\mathbf{R})$ are closed and bounded subsets of R, but $(-\infty, \infty)$ is not bounded.

This thorough examination of the Cantor set as a fractal produced by the iterated function system W provides the basis for similar results in higher dimensions. The principles remain unchanged and adding a dimension yields ever more interesting results.

2. Fractals in Two Dimensions

We now lift the setting from the real line to the Euclidean plane $\mathbf{R}^2$ and the space $(\mathcal{H}(\mathbf{R}^2), h(d))$ associated with it. Consider the following three contractions of the plane:

$$w_1\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$w_2\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/2 \\ 0 \end{bmatrix},$$

$$w_3\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/2 \end{bmatrix}.$$

Let $W = w_1 \cup w_2 \cup w_3$. Then W is a contraction on $\mathcal{H}(R^2)$, and has a unique fixed point, or attractor. To find the attractor, consider the action of W on the unit right triangle T with vertices $(0, 0)$, $(1, 0)$, and $(0, 1)$. Since T is a closed and bounded, nonempty subset of R^2, it is a point in $\mathcal{H}(R^2)$. Each of the maps w_1, w_2, and w_3, shrinks any point in R^2 towards the origin by a factor of $1/2$ in the directions of each coordinate axis. The map w_2 shifts each point $1/2$ unit to the right as well, while w_3 shifts the result $1/2$ unit upwards.

Figure 3.8 shows T and the result of $W(T)$. Notice that the result of $W(T)$ is to remove a "middle third" triangle from the original triangle T. Again, iterating T under W yields the results shown in Figure 3.9, displaying the remaining triangles from their previous iterations with the middle thirds removed.

Continuing to iterate W, we see that the fixed point of W is obtained by continually removing the middle third of each subtriangle from the original triangle T. This attractor is shown in Figure 3.10 and is called the **Sierpinski triangle** (or Sierpinski gasket).

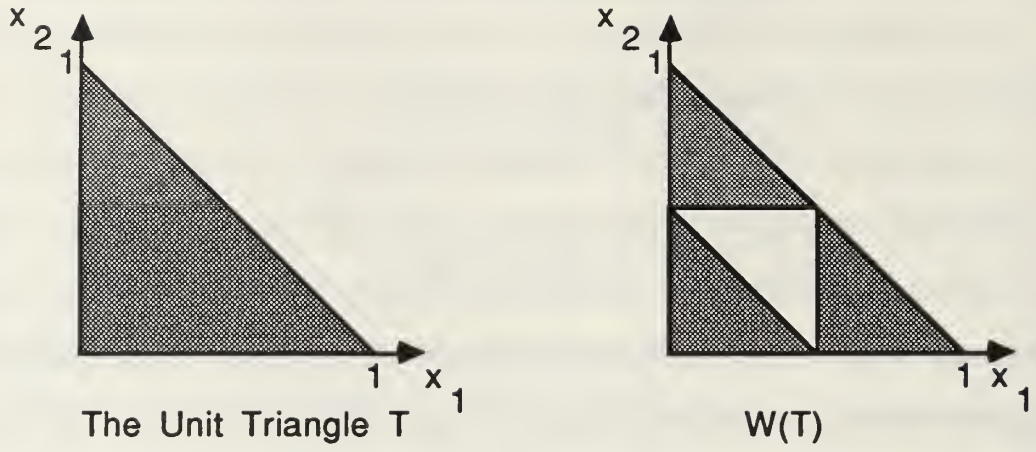

Figure 3.8 The effect of W on the unit triangle.

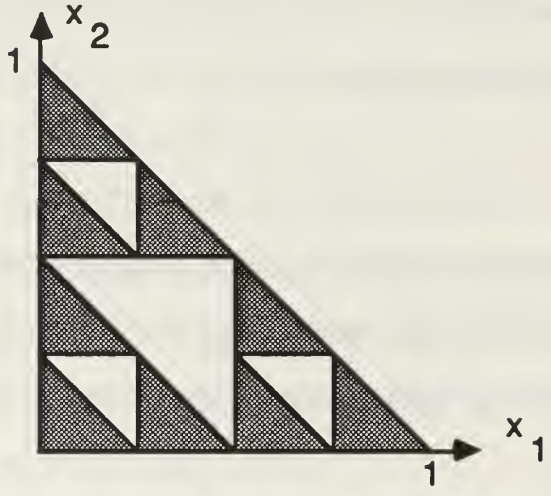

Figure 3.9 The effect of W^2 on the unit triangle.

The same results observed for the Cantor set in one dimension apply to the Sierpinski triangle in two dimensions. The Sierpinski triangle has infinite detail, and is self similar: each subtriangle contains a scaled copy of the original unit triangle. Furthermore, while we used the unit right triangle

T as the starting point, by the Contraction Mapping Theorem any point in the space $\mathcal{H}(\mathbb{R}^2)$ converges to the Sierpinski triangle under infinite iteration of W. Figure 3.11 shows how the unit square with infinite iteration under W also yields the Sierpinski triangle.

Figure 3.12 shows how the point (1, 1) starts away from the attractor, but under iteration of W eventually converges to points on the attractor. Figure 3.13 shows the effect of $W^6(S)$ on the unit square.

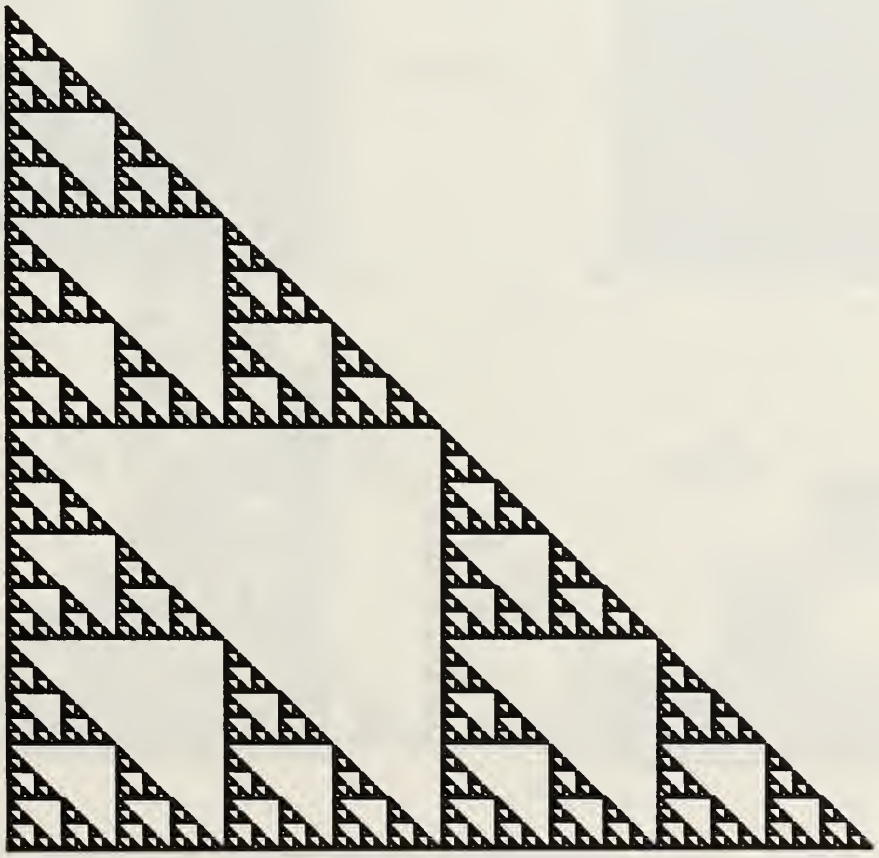

Figure 3.10 The Sierpinski triangle.

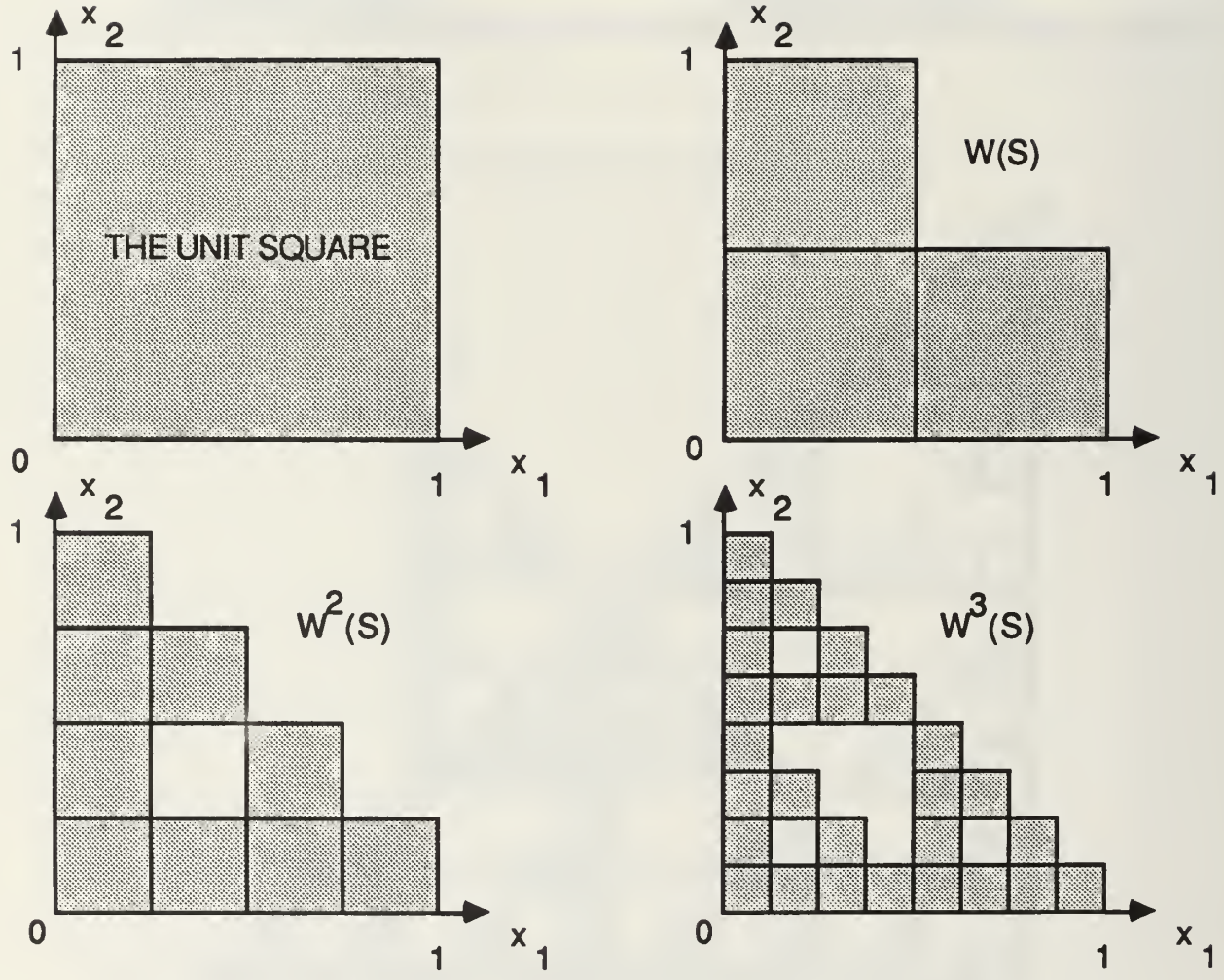

Figure 3.11 Creating the Sierpinski triangle by iterating the unit square.

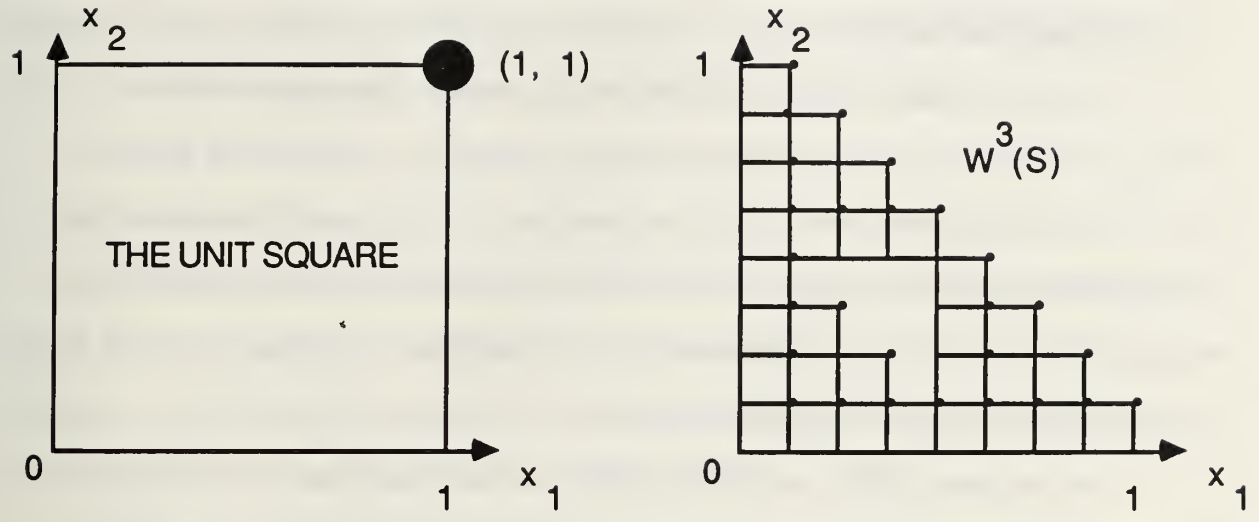

Figure 3.12 The orbit of the point (1, 1) under iteration of W.

Figure 3.13 The effect of $W^6(S)$ on the unit square.

3. Condensation Sets

So far we have only considered examples of contractions where $0 < s < 1$. We will now consider contractions where the contractivity factor is $s = 0$. Such contractions are called *condensations*, and in one dimension this corresponds to $f(x) = c$ where c is a constant. It is clear that no matter what two points we start with, the distance between their first iterates under f is 0, since $c - c = 0$, so that f is a contraction.

In the space $\mathcal{H}(X)$, $w_0: H(X) \to H(X)$ is a **condensation** if $w_0(x) = B$ for all x in $\mathcal{H}(X)$, where B is some fixed point in $\mathcal{H}(X)$. If we take a contraction mapping of the space $\mathcal{H}$, $W = w_1 \cup \ldots \cup w_n$, and form the union of it with w_0, where w_0 is a condensation such that $W' = W \cup w_0$, then W' will be a contraction with the same contractivity factor as W. To see this, consider that the contractivity factor of W is $s = \max\{s_1, s_2, \ldots, s_n\}$, and that the contractivity factor of W' is $s' = \max\{0, s_1, s_2, \ldots, s_n\} = s$.

Although the contractivity factors of the maps W and W' are the same, their dynamics are quite different. Since every iteration of W' maps points to the point B, the attractor of this iterated function system can be thought of as the infinite set of every iteration of B under W.

As an example, consider the one-dimensional case with the two functions $w_1(x) = x/2$ and $w_0(x) = 1$ on the real line. The map w_1 is a contraction, and w_0 is a condensation, and their fixed points are 0 and 1 respectively. If we start with a point on the real line, for example 0, and iterate it under $W = w_0 \cup w_1$, we see that $W(0) = \{0, 1\}$, $W^2(0) = \{0, 1, 1/2\}$, $W^3(0) = \{0, 1, 1/2, 1/4\}$, and $W^n(0) = \{0, 1, 1/2, 1/4, \ldots, 1/2^{n-1}\}$. Clearly, this

iterated function system generates an attractor which consists of every iterate of the condensation set $\{1\}$ under the contraction $w_1(x) = x/2$.

4. A Fractal Tree

We now use the concept of condensation sets to construct a fractal image. We first let the set $B = \{(x_1, x_2): -0.1 \le x_1 \le 0.1; 0 \le x_2 \le 1\}$. In other words, B is the filled rectangle with height 1 and width .2 with base centered at the origin. Now define the transformations

$$w_0\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = B,$$

$$w_1\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} .75\cos(\pi/4) & -.75\sin(\pi/4) \\ .75\sin(\pi/4) & .75\cos(\pi/4) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$w_2\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} .75\cos(-\pi/4) & -.75\sin(-\pi/4) \\ .75\sin(-\pi/4) & .75\cos(-\pi/4) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The transformations w_1 and w_2 shrink any image by three-fourths in the direction of both axes, w_1 rotates it 45° clockwise and w_2 rotates it 45° counterclockwise, and then both transformations shift the image up one unit. The transformation w_0 is a condensation that maps any point to the set B.

If we let $W = w_0 \cup w_1 \cup w_2$, and we iterate the origin under W, we see that $W(0, 0)$ yields the set B, W^2 yields the set B together with the two condensed, rotated, shifted copies of B, and W^3 yields the set B together with two condensed, rotated, and shifted copies of the image of W^2. If we continue iterating, we see that every iteration yields the set B together with

two condensed, rotated, and shifted copies of the previous iteration. The first seven iterations of the origin under W are shown in Figure 3.14, which is the beginning of a fractal "tree."

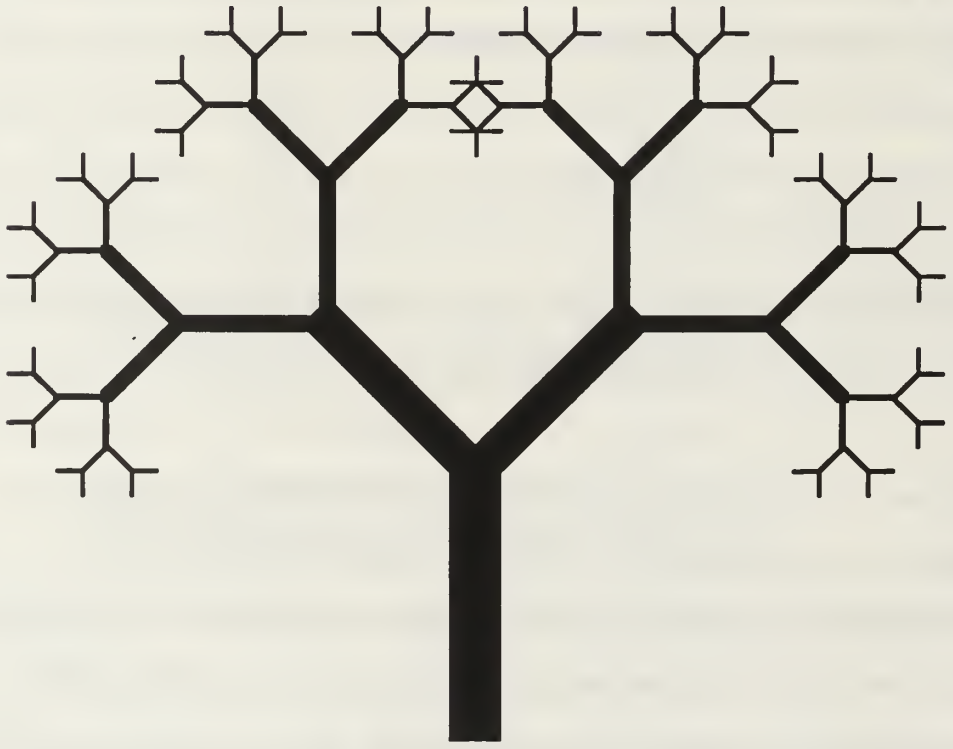

Figure 3.14 A fractal tree.

Another way of generating a fractal tree is through an iterated function system without condensation. If we consider the transformations

$$w_1\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}.2 & 0\\0 & .5\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}+\begin{bmatrix}.45\\0\end{bmatrix},$$

$$w_2\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}.1\cos(\pi/4) & -.25\sin(\pi/4)\\.1\sin(\pi/4) & .25\cos(\pi/4)\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}+\begin{bmatrix}.5\\.5\end{bmatrix},$$

$$w_3\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}.1\cos(-\pi/4) & -.25\sin(-\pi/4)\\.1\sin(-\pi/4) & .25\cos(-\pi/4)\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}+\begin{bmatrix}.5\\.5\end{bmatrix},$$

and we consider the effect of $W = w_1 \cup w_2 \cup w_3$ on the unit square, we see the result shown in Figure 3.15. If we continue to iterate W, we produce the attractor of this iterated function system, which is a fractal tree with a similar shape to the one in Figure 3.14. However, unlike the tree with condensation, the trunk and branches of this tree are not solid, in fact, each contains a condensed copy of the entire attractor. These detailed images are not only aesthetically pleasing, but they lead to an important application of fractals. To store these images in a computer using traditional means would require hundreds of lines of code; however, we have used just 18 numbers (three two-by-two matrices and three two-vectors) to record all of the information required to produce these images. This efficient way of recording information has led to important applications of fractal geometry, as the next section shows.

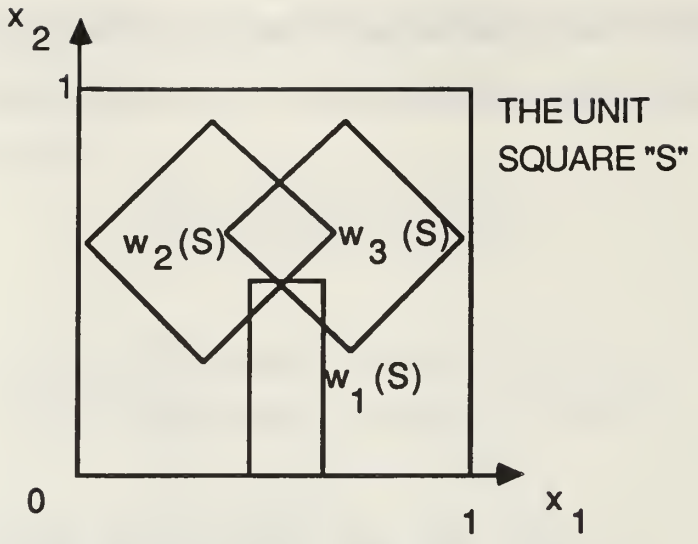

Figure 3.15 The action of W on the unit square.

F. APPLICATIONS OF FRACTALS TO COMPUTER GRAPHICS

Pictures that are infinitely detailed and self similar can be stored in a computer with very small amounts of information. For example, in the case of the fractal tree in the last section, only 18 numbers were required. But what if an image is not self similar? Is it still possible to use contraction mappings to store the information? The remarkable answer is "yes." Perhaps even more surprising, this procedure can be achieved to any desired degree of accuracy.

A formal description of using contractive iterated function systems to store graphical information can be found in Barnsley (1988) or Falconer (1990). Instead of presenting the proofs here, we give an intuitive treatment of the technique through an example. Then we show how any desired degree of accuracy can be attained.

Recall that affine transformations can have many different actions on a set, including shearing, shrinking or stretching, rotating, and shifting. Suppose we have an image we wish to reproduce; let that be our attractor. We then search for affine transformations that could paste distorted copies of the set back on top of itself to approximate the original image as closely as possible. This is done through the "collage theorem" due to Barnsley. We present an example here.

Consider the image in Figure 3.16. While this figure is not self similar, we can approximate it by making smaller copies of it (through contractions) and "pasting" them back on itself (through affine transformations) to cover,

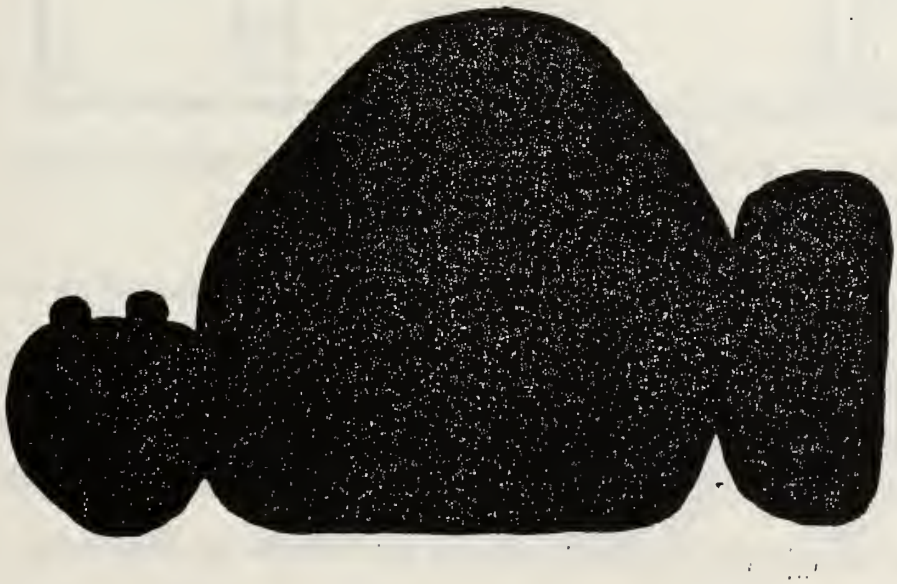

Figure 3.16 A nonself-similar image.

as much as possible, the original figure. One such covering is shown in Figure 3.17. The four corners of the original image and their movement under the coverings are indicated by the letters A through D. Note that a certain amount of overlapping of the covering is required, for if this were just-touching or totally disconnected, the attractor would be too sparse to represent the original image.

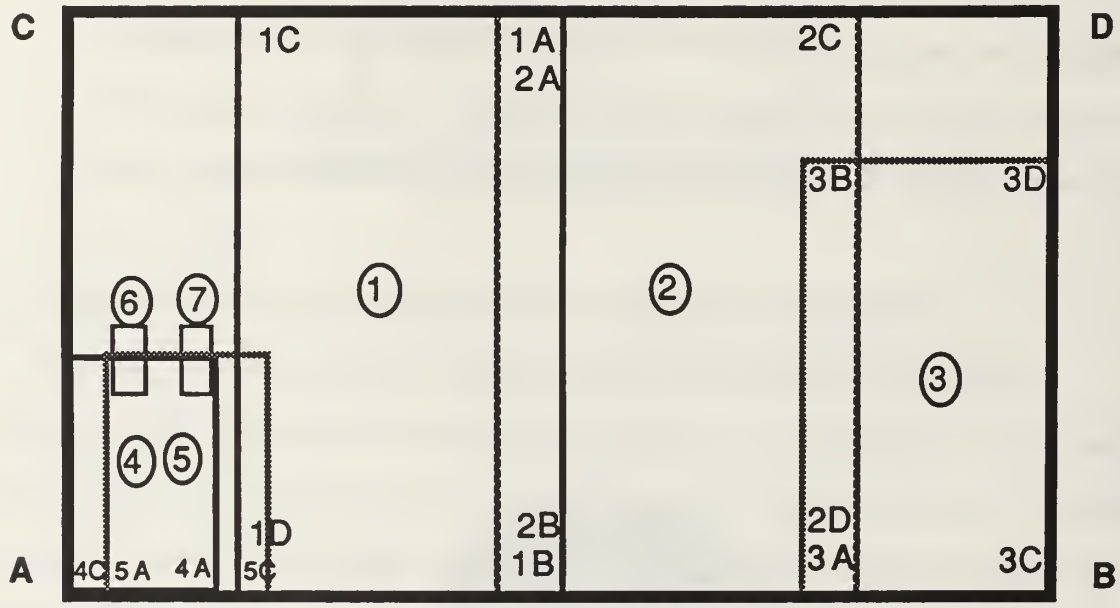

Figure 3.17 Covering an image with transformed copies of itself.

Recalling that an affine transformation is uniquely determined by its action on any three noncollinear points, we could easily determine the actual transformations required to produce the seven mappings indicated in the figure. Under repeated iteration of this iterated function system, we could then reproduce the image in a similar way as that used to produce other attractors, with only 42 pieces of data, since each affine transformation requires six numbers.

But notice that the attractor of this iterated function system is not the original image, since its boundaries are infinitely detailed, as opposed to the solid boundaries in the original image (see Figure 3.18). This "fuzziness" seems to contradict the earlier claim that we could approximate a computer image to any degree of accuracy. But the above system used only seven transformations. Noting that any computer image is displayed by a finite number of pixels, which are either on or off, we could push this technique to its extreme to reproduce exactly the original image. That is, we could condense the entire set down to the size of one pixel, and then shift that pixel through affine transformations to every pixel in the original image. This will indeed give an exact reproduction of the original image. However, it would require six numbers for every pixel in the image.

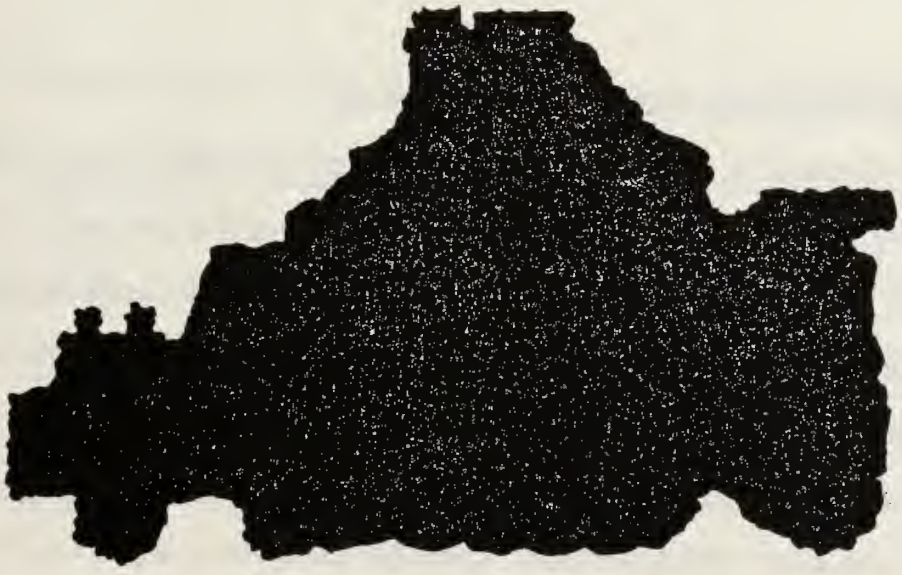

Figure 3.18 The attractor with a fractal-like boundary.

While this technique obviously defeats the concept of efficient storage of information, it nevertheless shows that we can achieve any desired degree of accuracy. Depending on the requirements of a given project, iterated function systems and their attractors can frequently be used to achieve the desired degree of accuracy much more efficiently than the traditional methods of storing information.

So far we have claimed that only the information contained in the affine transformations is needed to use this technique. In fact, we also require a program to produce the image based on the input. However, these are quite common and there are a variety of algorithms that produce the desired results. Thus, in a situation where many different figures must be recorded, a single program will suffice for creating the graphics once the transformations are stored. It follows that this method is more efficient than many traditional computer graphics techniques.

G. THE ADDRESSES OF POINTS ON FRACTALS

A useful technique in analyzing fractals is to address every point on an attractor by the sequence of transformations that led to that point being in the attractor. This leads to the classification of iterated function systems as being "totally disconnected," "just touching," or "overlapping," and helps us analyze chaotic dynamics on fractals in Chapter IV.

We will start with transformations of the real line to illustrate this concept in the simplest setting possible. Consider the system $w_0(x) = x/3$ and $w_2(x) = x/3 + 2/3$ (the reason for selecting these subscripts will soon become apparent). We already know that repeated iteration of $W = w_0 \cup w_2$ yields

the Cantor middle-thirds set as an attractor. (Recall that there are no intervals in the Cantor set, and that the intersection of w_0 and w_2 applied to the unit interval, or any subset thereof, is empty). Considering the iteration of W on the unit interval, we now begin to address the points in its attractor.

Under the first iteration of W, the unit interval is mapped to the interval [0, 1/3] by w_0 and to [2/3, 1] by w_2. Hence, we begin the address of every point in [0, 1/3] with 0, and we begin the address of every point in [2/3, 1] with 2. Under the second iteration of W, the interval [0, 1/3] is mapped to [0, 1/9] by w_0 and to [2/3 ,7/9] by w_2, whereas the interval [2/3, 1] is mapped to [2/9, 1/3] by w_0 and to [8/9, 1] by w_2. Hence, the second number in the address of all points in [0, 1/9] and [2/3 , 7/9] is 0, and the second number in the address of all points in [2/9, 1/3] and [8/9, 1] is 2. Figure 3.19 shows the first three steps of this process, and the beginning of the resulting addresses of every point in the Cantor set.

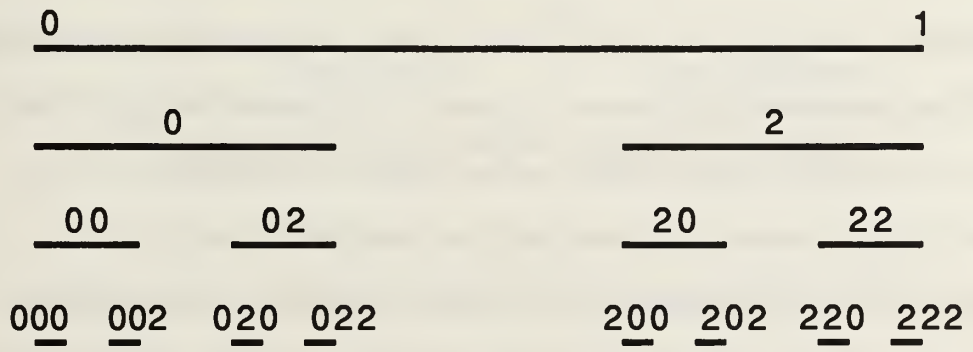

Figure 3.19 Addressing the attractor of an iterated function system.

It is clear that under infinite iteration, every point in the Cantor set has a unique address consisting of an infinite string of 0s and 2s. Note that the

origin has the address 000..., and that 1 has the address 222.... What we have by the convenient selection of 0 and 2 as the subscripts is that every point in the Cantor set is addressed by its ternary expansion, as described in Chapter II.

The addressing scheme used here is not unique, as we could have also chosen 0 and 1 as subscripts resulting in every point in the Cantor set having an address from the code space Σ_2. Again, note that every point in Σ_2 would be an address of a point in the Cantor set, and that any dynamics we performed on the code space Σ_2 could easily be applied to the Cantor set. This one-to-one correspondence between code space and fractals is very useful in the discussion of chaotic dynamics on attractors of iterated function systems. (As an aside, we could have just as easily selected 1 and 2 as subscripts to demonstrate further that an addressing scheme is not unique. Nevertheless, all of these systems have a one-to-one correspondence between them.)

Notice that every point in the Cantor set has a unique address associated with it. This characteristic allows us to classify the iterated function system $W = w_0 \cup w_2$ as being **totally disconnected**. While the attracting set (the Cantor set) is disconnected, we associate the classification "totally disconnected" not with the attractor, but rather with the iterated function system that produced it. The reason for this will become apparent when we look at "just touching" and "overlapping" iterated function systems which have the same set as their attractor.

Now consider the iterated function system $W = w_1 \cup w_2$, where $w_1(x) = x/2$ and $w_2(x) = x/2 + 1/2$. The unique fixed point (and hence the

attractor) of this system is clearly the unit interval. So if we consider the action of this system on [0, 1], we can build addresses for every point in that interval. Starting with [0 ,1], we see that w_1 maps it to [0, 1/2] so we assign to that interval the first address of 1, and similarly points in [1/2, 1] have the first address 2. Already we can see that there is some ambiguity with the point 1/2, since it appears to be receiving two distinct addresses. Continuing, we see that points in the interval [0, 1/4] have addresses starting with 11, while points in the intervals [1/4, 1/2], [1/2, 3/4], and [3/4, 1] have addresses 12, 21, and 22, respectively. Now the points 1/4 and 3/4 have been added to the list of points with dual representations. Figure 3.20 shows the first three iterations of this addressing scheme.

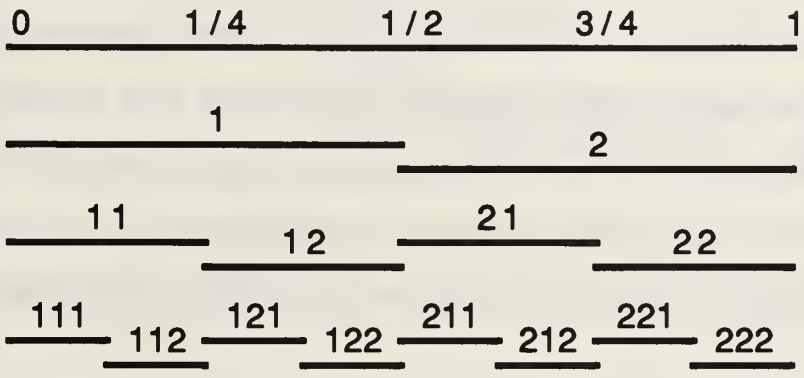

Figure 3.20 Addressing the attractor of a just-touching IFS.

Again, it is clear that every point in the unit interval receives an address, but points of the form $i/2^n$, $i = 1, 2, 3, \ldots, 2^n - 1$ have dual representations under this scheme. The set of points where the addresses touch, although infinite, is **countable** (i.e., is in one-to-one correspondence with the natural

numbers) since at every step we add a finite number of points with dual addresses. Also, notice that no intervals have dual addresses, since between every pair of points with dual addresses, there are points with unique addresses. Iterated function systems with this characteristic are classified as **just-touching**. This fits the intuitive concept of this definition since the results of iterating the attractor [0, 1] under w_1 and w_2 results in two intervals which "just touch" at one point.

Finally, we consider the system $W = w_1 \cup w_2$ where $w_1(x) = 2x/3$ and $w_2(x) = 2x/3 + 1/3$. Again, the attractor of this system is clearly [0, 1], but when we address points in the attractor, we encounter considerable ambiguity. Under the first iteration of W applied to [0, 1], we see that the interval [0, 2/3] gets the address 1, while the interval [1/3, 1] gets the address 2. Here an entire interval [1/3, 2/3] of points has a dual address after just the first iteration. Moreover, at every step we add intervals with a similar ambiguity. The first three steps in this addressing scheme are shown in Figure 3.21. Notice that in order to remove the ambiguity of the overlapping addresses, we have to "lift" the unit interval into a second dimension as the figure shows.

The ambiguity in this addressing system is much greater than for the just-touching iterated function system. In fact, at every step an uncountable number of points with multiple addresses is obtained. Here, the number of points with multiple addresses far outnumbers the points with unique addresses. What is more, for this system, the only points with unique addresses are 0 and 1. This feature leads to the intuitive definition of an **overlapping** iterated function system. Notice that each iteration of w_1 and

w_2 results in an interval where different addresses overlap. Because the attractor of this system is the same as in the example of a just-touching iterated function system, it should now be clear why the classifications "totally disconnected," "just touching," and "overlapping" are applied to the iterated function system rather then to the attractor itself.

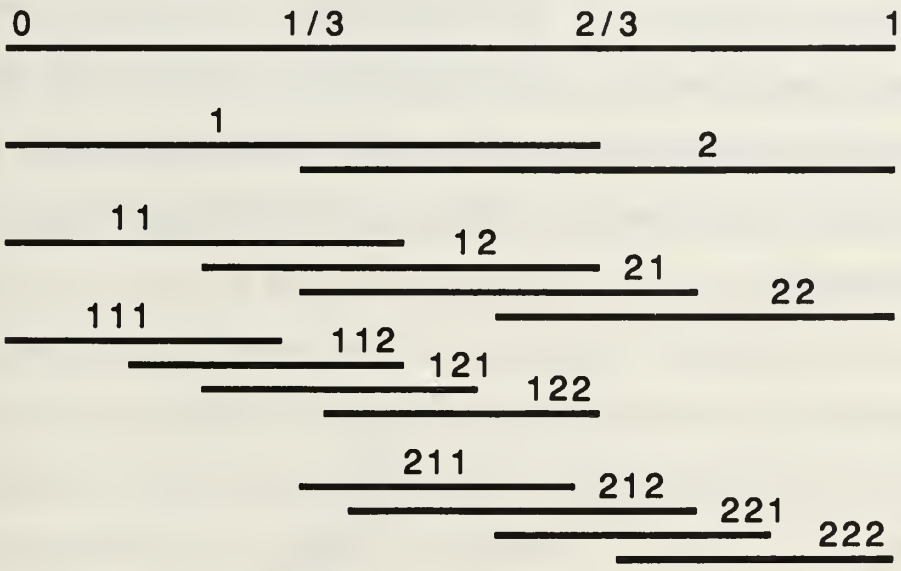

Figure 3.21 Addressing the attractor of an overlapping IFS.

With the concept of addressing points on attractors and classifying iterated function systems in a simple one-dimensional setting, let us now turn to a two- dimensional example. Recall the iterated function system that produced the Sierpinski triangle as an attractor:

$$w_1\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$w_2\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/2 \\ 0 \end{bmatrix},$$

$$w_3\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/2 \end{bmatrix},$$

where $W = w_1 \cup w_2 \cup w_3$. Iterating the unit right triangle under W yields the Sierpinski triangle by removing the "middle-thirds triangle" at every step. Looking at just one iteration of W, we see that this is a just-touching iterated function system: the image of w_1 intersects the image of w_2 at (1/2, 0), the images of w_2 and w_3 intersect at (1/2, 1/2), and the images of w_1 and w_3 intersect at (0, 1/2). An addressing scheme for the Sierpinski triangle is shown in Figure 3.22.

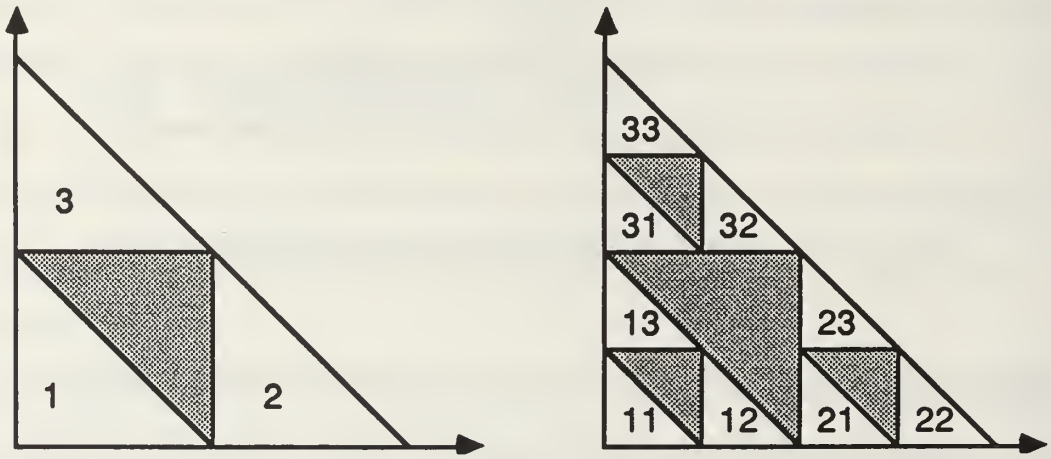

Figure 3.22 An addressing scheme for the Sierpinski triangle.

As before, the vertices created at each step have two distinct addresses. For example, the point (1/2, 1/2) has the addresses 2333. . . and 3222. . . . The exceptions are the three original vertices which have the unique

addresses 111..., 222..., and 333.... Again, we could have chosen the addresses to consist of 0s, 1s, and 2s, to create a correspondence between the Sierpinski triangle and code space Σ_3. However, it is just as convenient to create a modified code space Σ'_3 to be $\{x_1x_2x_3\ldots : x_i \in \{1, 2, 3\}\}$. The metric associated with this space Σ'_3 is

$$d(x, y) = \sum_{i=1}^{\infty} \frac{|x_i - y_i|}{4^i},$$

so Σ'_3 is clearly in one-to-one correspondence with the Sierpinski triangle.

A more difficult addressing problem occurs with an overlapping iterated function system in two dimensions. Recall that in an overlapping iterated function system of the real line, we had to "lift" the attractor into a second dimension to uniquely determine addresses. Similarly, in two dimensions, we must lift the attractor into a third dimension to address its points uniquely. For example, consider the overlapping iterated function system of the plane $W = w_1 \cup w_2 \cup w_3$, where

$$w_1\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2/3 & 0 \\ 0 & 2/3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$w_2\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2/3 & 0 \\ 0 & 2/3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/3 \\ 0 \end{bmatrix},$$

$$w_3\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2/3 & 0 \\ 0 & 2/3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/3 \end{bmatrix}.$$

The attractor for this iterated function system is the "filled" Sierpinski triangle, or simply the unit right triangle. Figure 3.23 shows the first iteration of W and how the iterates of w_1, w_2, and w_3 overlap. Any

point along the x_1-axis in the interval $[1/3, 2/3]$ has an address beginning with a 1 or a 2, and entire regions of the plane have similar ambiguities.

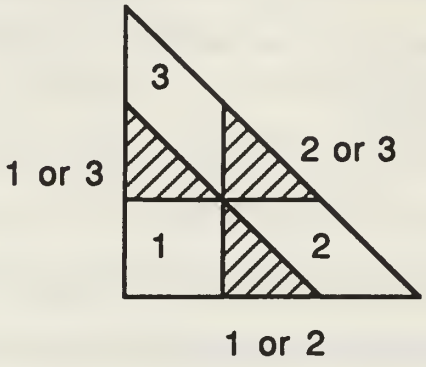

Figure 3.23 Ambiguous addresses for the attractor of an overlapping IFS.

While points on the attractor always correspond to multiple addresses, we would like to identify every possible address with a unique point. This is useful, for example, when considering the dynamics of the shift map on code space associated with the attractor of an iterated function system. Developing such an addressing scheme in two dimensions for this attractor would be extremely difficult. However, "lifting" the attractor into a third dimension allows us to create an unambiguous address system for the attractor. Figure 3.24 shows how the first iterate of W (from Figure 3.23) is lifted into the third dimension. This removes the ambiguous regions from Figure 3.23.

We now lift the entire attractor into the third dimension. To accomplish this, cross the Euclidean plane with the unit interval $[0, 1]$ which is represented by its base four expansion so that every point in the interval is represented by a number $.x_1x_2x_3\ldots$ where $x_i \in \{0, 1, 2, 3\}$. By associating the

address of each iterate of W with its base four expansion, we see that we get a Cantor-type set construction along the vertical axis when viewed from the side. Since we used the subscripts 1, 2, and 3 for the iterated function system, every number in [0, 1] that does not have a 0 in its base four expansion will have an iterate associated with it. The result of this process, taken to three iterates, is shown in figure 3.25. Having an addressing scheme such as this to locate a unique point proves quite useful when we study chaotic dynamical systems on fractals in Chapter IV.

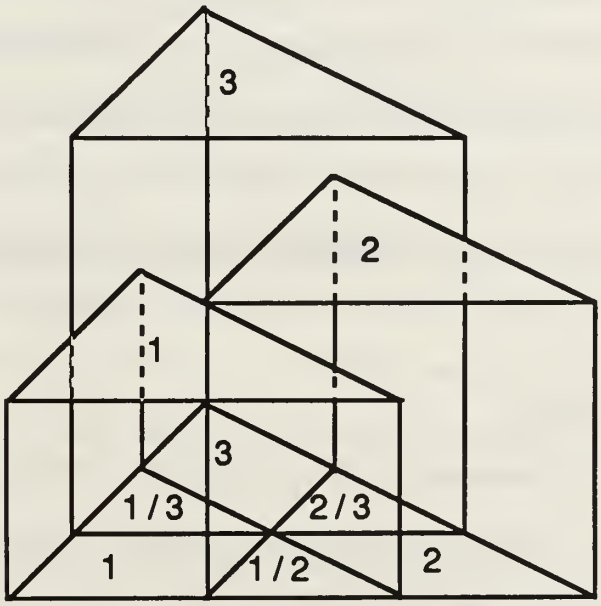

Figure 3.24 Lifting an attractor into the third dimension.

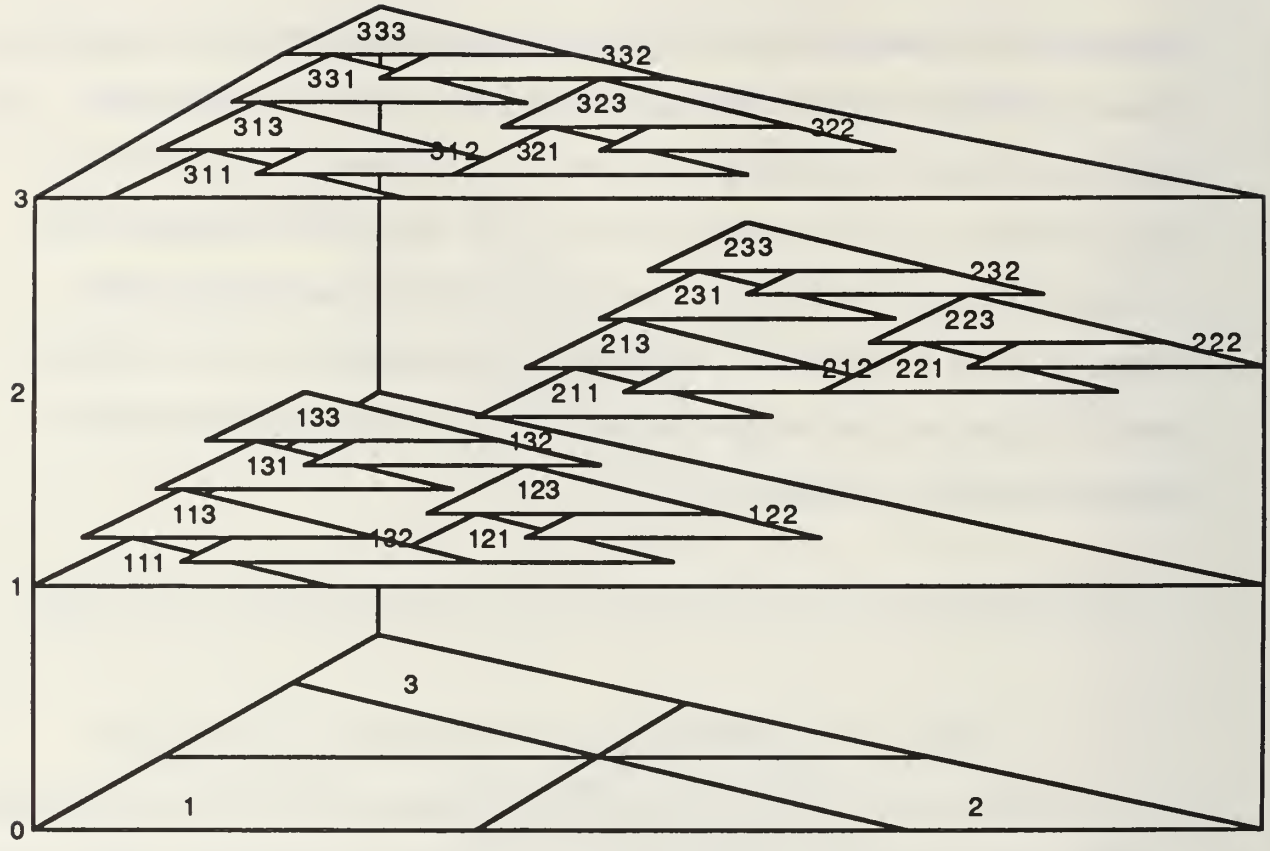

Figure 3.25 Addressing the lifted attractor of an overlapping IFS.

H. FRACTAL DIMENSION

One of the most useful concepts in the application of fractal geometry is that of fractal dimension. Fractal dimension provides a measure of the size or "dimension" of an object, whether it is the attractor of an iterated function system, a more familiar geometric figure, or something arrived at through collection of real-world data. Fractal dimension has worked its way into many fields such as physics, meteorology, aeronautical engineering, and

oceanography, and we provide some examples of these applications shortly. To begin we develop this concept mathematically.

The fractal dimension of an object is the assignment of a number to the object that represents how much space the object takes up in its ambient space. We are all familiar with objects in one, two, or three dimensions. That an object may have a non-integer or fractional dimension seems counterintuitive. Nevertheless, you will see that the definition of fractal dimension supports completely our intuitive concept of one, two or three dimensional objects. With this in mind, we give some preliminary definitions that are needed to develop the concept of fractal dimension.

An **ε-ball** about a point x_0 in a metric space is the set of all points in the space within distance $\varepsilon > 0$ from the point x_0. Conventional notation for an ε-ball is $B(x_0, \varepsilon) = \{x: d(x, x_0) \leq \varepsilon\}$. Notice that by using the "≤" sign in the definition, the ε-balls in our discussion are closed. While this notation may be slightly unfamiliar, the concept certainly is not. For if we take the real line with the standard distance function as our metric space, then $B(1, 1/2)$ is simply the closed interval $[0.5, 1.5]$. Furthermore, if the metric space is the Euclidean plane R^2, then $B((0, 0),1)$ is simply the closed unit disc centered at the origin.

Given a closed bounded nonempty subset of a metric space, we want to **cover** that subset with balls of specified radius ε. In the case of the real line, to cover the unit interval with balls of radius $1/4$ requires at least two (centered at $1/4$ and $3/4$). We could use any number of such balls if we allow overlapping. We are primarily interested in covering a set with the smallest number of balls possible, so we define the integer $N(A, \varepsilon)$ to be the smallest

number of closed balls of radius ε needed to cover the set A. Specifically, for any set A in our metric space, we define $N(A, \varepsilon)$ to be the smallest positive integer M such that

$$\bigcup_{n=1}^{M} B(x_n, \varepsilon) \supset A,$$

for some distinct set of points $\{x_n: n = 1, 2, \ldots, M\}$ in the metric space.

The Heine-Borel Theorem (see an advanced calculus text) guarantees that we are always able to cover a closed bounded set with a finite number of ε-balls. Since every set of positive integers has a smallest member, the number $N(A, \varepsilon)$ uniquely exists.

We are now prepared to define fractal dimension. If we let A be a point in the space $\mathcal{H}(X)$ (i.e., A is a closed bounded nonempty subset of the metric space X) the quantity D is defined as

$$D = \lim_{\varepsilon \to 0} \frac{\ln(N(A, \varepsilon))}{\ln(1/\varepsilon)}.$$

The number D, if it exists, is called the **fractal dimension** of A and is denoted $D(A)$.

Let us clarify this concept with some intuitive examples. First, consider a point x in R^2. No matter how small we choose ε, we can always cover the point x with a single ball of radius ε, so that $D(x) = \lim_{\varepsilon \to 0} \ln(N(x, \varepsilon))/\ln(1/\varepsilon) = \lim_{\varepsilon \to 0} \ln(1)/\ln(1/\varepsilon) = 0$. Hence, for any point x in R^2, we see that its fractal dimension is zero (which fits our intuitive concept of a point having zero dimension).

Consider next the unit interval $[0, 1]$ as a subset of the real line. Recalling that in R, ε-balls are the closed intervals of length 2ε, we begin to

cover [0, 1] with ε-balls. If we let $\varepsilon = 1/2, 1/4, 1/8, \ldots, 1/2^n, \ldots$ then the number of intervals required to cover [0, 1] are 1, 2, 4, . . . , and 2^{n-1}, respectively. Clearly, $n \to \infty$ implies $\varepsilon \to 0$, so $D([0, 1]) = \lim_{\varepsilon \to 0} \ln(N(A, \varepsilon))/\ln(1/\varepsilon) = \lim_{n \to \infty} \ln(2^{n-1})/\ln(2^n) = \lim_{n \to \infty} (n-1)\ln 2/n \ln 2 = \lim_{n \to \infty} (n - 1)/n = 1$. Hence, the fractal dimension of a line segment is 1, which is consistent with our ordinary concept of dimension.

In the previous example we were very specific in the way we let $\varepsilon \to 0$. This example leads us to an equivalent definition of fractal dimension. In the same setting as before, let $\varepsilon_n = r^n$ for $0 < r < 1$ and $n = 1, 2, 3, \ldots$ Then

$$D(A) = \lim_{n \to \infty} \ln(N(A, \varepsilon_n))/\ln(1/\varepsilon_n).$$

To show that these definitions are equivalent, let $f(\varepsilon) = \text{Max}\{\varepsilon_n: \varepsilon_n < \varepsilon\}$, and assume that $\varepsilon \le r$. Then $f(\varepsilon) \le \varepsilon \le f(\varepsilon)/r$, and $N(A, f(\varepsilon)) \ge N(A, \varepsilon) \ge N(A, f(\varepsilon)/r)$. Since, for $x > 1$, $\ln(x)$ is a positive increasing function, we have

$$\frac{\ln(N(A, f(\varepsilon)/r))}{\ln(1/f(\varepsilon))} \le \frac{\ln(N(A, \varepsilon))}{\ln(1/\varepsilon)} \le \frac{\ln(N(A, f(\varepsilon)))}{\ln(r/f(\varepsilon))}.$$

We can assume that $N(A, \varepsilon) \to \infty$ as $\varepsilon \to 0$, or the result follows immediately. Considering the left side, and taking the limit as $\varepsilon \to 0$, since for some n, $f(\varepsilon) = \varepsilon^n$, we have

$$\lim_{\varepsilon \to 0}\left[\frac{\ln(N(A, f(\varepsilon)/r))}{\ln(1/f(\varepsilon))}\right] = \lim_{n \to \infty}\left[\frac{\ln(N(A, \varepsilon_{n-1}))}{\ln(1/\varepsilon_n)}\right] =$$

$$\lim_{n \to \infty}\left[\frac{\ln(N(A, \varepsilon_{n-1}))}{\ln(1/r) + \ln(1/\varepsilon_{n-1})}\right] = \lim_{n \to \infty}\left[\frac{\ln(N(A, \varepsilon_n))}{\ln(1/\varepsilon_n)}\right].$$

Similarly, the right side yields

$$\lim_{\varepsilon \to 0}\left[\frac{\ln(N(A, f(\varepsilon)))}{\ln(r/f(\varepsilon))}\right] = \lim_{n \to \infty}\left[\frac{\ln(N(A, \varepsilon_n))}{\ln(r/\varepsilon_n)}\right] =$$

$$\lim_{n\to\infty}\left[\frac{\ln(N(A,\varepsilon_n))}{\ln(r)+\ln(1/\varepsilon_n)}\right]=\lim_{n\to\infty}\left[\frac{\ln(N(A,\varepsilon_n))}{\ln(1/\varepsilon_n)}\right].$$

The pinching theorem from calculus establishes the equivalence of these definitions.

Notice that in R^2 with the Euclidean metric, the ε-balls always give closed discs. Frequently we find it more convenient to use "boxes" or closed squares to cover our sets. For example, if we want to cover the unit square [0, 1]X[0, 1], it would be much easier to calculate the required number of boxes of a given size than to calculate the number of ε-balls. It turns out that we can obtain the same value of fractal dimension using closed boxes as we obtain using ε-balls.

First, consider a grid in R^2. Figure 3.26 shows a covering of this grid by closed discs, and we see that if we were to refine our grid and let the radii of the discs go to zero, then the number of boxes and the number of discs will remain in one-to-one correspondence.

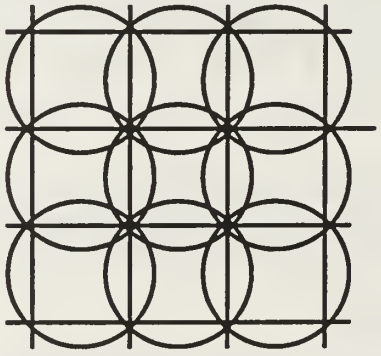

Figure 3.26 Covering a grid with closed discs.

The result of placing finer and finer grids with side length $1/2^n$ over a set in R^2, and letting $N_n(A)$ be the number of grid squares that intersect the object, gives a third equivalent definition of fractal dimension as follows:

$$D(A) = \lim_{n\to\infty}\left[\frac{\ln(N_n(A))}{\ln(2^n)}\right].$$

Actually several equivalent definitions of fractal dimension are used in different applications, depending on which is the most convenient. All of the definitions give the same result. Therefore, both theoretical and experimental applications are quite easy to perform in many different situations. Different examples of these definitions for Euclidean spaces are

$$D(A) = \lim_{\varepsilon\to 0}\left[\frac{\ln(N(A,\varepsilon))}{\ln(1/\varepsilon)}\right],$$

where $N(A, \varepsilon)$ is

1. The smallest number of closed balls of radius ε that cover A (our first definition);
2. The smallest number of boxes of side ε that cover A;
3. The smallest number of sets of diameter of at most ε that cover A (where the diameter of a general set is the largest distance between all pairs of points in the set); and
4. The largest number of disjoint balls of radius ε with centers in A.

We have presented these definitions in the context of R and R^2, but they are easily extended to R^3, where ε-balls become closed spheres and ε-boxes become cubes. More generally, the same theory holds for R^n for any integer n, although the applications are of a more theoretical nature.

Now that we have discussed equivalent definitions of fractal dimension, let us use the most convenient one to compute the dimension of the closed unit square [0, 1]X[0, 1]. Using squares with side length $1, 1/2, 1/4, \ldots 1/2^n$ in R^2, we see that $1, 4, 16, \ldots, 4^n$ of them are required to cover the unit square, respectively. Then

$$D([0, 1]X[0, 1]) = \lim_{n\to\infty}\left[\frac{\ln(N_n(A))}{\ln(2^n)}\right],$$

we have

$$D([0, 1]X[0, 1]) = \lim_{n\to\infty}\left[\frac{\ln(4^n)}{\ln(2^n)}\right] = \lim_{n\to\infty}\frac{n\ln 4}{n\ln 2} = 2.$$

Hence, our definition of fractal dimension again supports our knowledge that the closed unit square has dimension 2.

We now apply the definition of fractal dimension to the Cantor set and the Sierpinski triangle. Intuitively, one might believe that, since these objects lie in R and R^2 respectively, they should have dimensions 1 and 2. But recall that we removed a certain amount of length or area at each iteration for an infinite number of iterations. Thus it might be plausible for these objects to have a smaller dimension than that of the spaces in which they exist.

Consider first the Cantor set as a subset of the unit interval. Let the ε-balls in R be the intervals of length $1, 1/3, 1/9, \ldots, 1/3^n$. One can easily see that $1, 2, 4, \ldots, 2^n$ balls are required to cover the set, respectively. Hence, the fractal dimension of the Cantor set is

$$D(C) = \lim_{n\to\infty}\left[\frac{\ln(2^n)}{\ln(3^n)}\right] = \lim_{n\to\infty}\frac{n\ln 2}{n\ln 3} = \frac{\ln 2}{\ln 3},$$

so that the Cantor set has dimension of approximately .6309. We removed a total length of 1 from the unit interval when constructing the Cantor set, but still left an uncountable number of points. Thus intuitively we might expect its dimension to be somewhere between 0 and 1. Fractal dimension provides a measure of how many points actually remain in the set.

Moving next to the Sierpinski triangle, if we use squares of side length 1, $1/2, 1/4, \ldots, 1/2^n$, then $1, 3, 9, \ldots, 3^n$ squares are required to cover the set (see Figure 3.27).

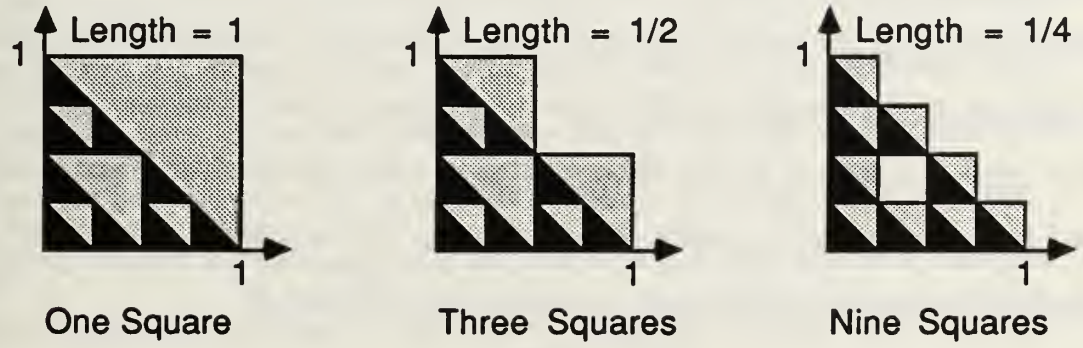

Figure 3.27 Covering the Sierpinski triangle with grids of size $1/2^n$.

Hence, the dimension of the Sierpinski triangle is given by

$$D(s) = \lim_{n\to\infty}\left[\frac{\ln(3^n)}{\ln(2^n)}\right] = \lim_{n\to\infty}\frac{n\ln 3}{n\ln 2} = \frac{\ln 3}{\ln 2},$$

and the Sierpinski triangle has dimension of approximately 1.585. Again, considering the amount of area removed from the unit triangle, it seems reasonable that the dimension should be between 1 and 2. It now becomes clear why the term "fractal" was chosen to describe objects like the Cantor set

and the Sierpinski triangle, as they have dimensions which are fractions of whole numbers.

While we have shown how to determine the fractal dimension theoretically in some very simple cases, we are not always able to make this computation by just looking at the attractor geometrically. However, if the iterated function system that created the attractor is known, the fractal dimension of the attractor can be determined by analyzing the iterated function system itself. We discuss this idea next, and state a general result.

If we know that $W = w_1 \cup w_2 \cup \ldots \cup w_n$ is the iterated function system producing an attractor, and if each w_i is a similitude with contractivity factor s_i (as defined in Section III.3, in that it is a shrinking in each direction by the factor r such that $s_i = r_i$, followed by a rotation and a translation) then if the iterated function system is totally disconnected or just touching, the attractor has fractal dimension $D(A)$ given by the unique solution of

$$\sum_{i=1}^{n} |s_i|^{D(A)} = 1,$$

where $D(A)$ will be between 0 and the dimension of the ambient Euclidean space. If the iterated function system is overlapping, then we only get a bound for $D(A)$ given by $D(A) \leq D$, where

$$\sum_{i=1}^{n} |s_i|^{D} = 1.$$

Even if we do not know the exact iterated function system that produced the attractor, we can measure the contraction of our set under each similitude, and solve for a rough approximation of $D(A)$. While the proof of this result is quite complicated, (see Barnsley, 1988, p. 185) it does give a very useful theoretical tool for determining the fractal dimension of an attractor.

To demonstrate its use on the Sierpinski triangle, which we know is created with three similitudes with contractivity factor of 1/2, and knowing that the Sierpinski triangle is just touching, we have $(1/2)^D + (1/2)^D + (1/2)^D = 1$, from which we get $D = \ln 3/\ln 2$, as expected. Unfortunately, the theoretical determination of fractal dimension is not always possible, so we must frequently use experimental methods in applications.

I. EXPERIMENTAL DETERMINATION OF FRACTAL DIMENSION

Fractal dimension, as previously discussed, is perhaps the most useful aspect of fractal geometry in real-world applications. It has permitted the analysis of numerous natural phenomena that previously were inaccessible to scientists because of a lack of tools and theory. In nature, nice standard geometric shapes are primarily the exception rather than the rule; irregular, fractal-like shapes are found everywhere. Consider, for instance, the shapes of clouds, coastlines, geological formations, atmospheric phenomena, plants, crystals, and microorganisms. None of these objects can be described by a "nice" geometric shape, but fractal geometry has provided a new way of analyzing them.

One way scientists have studied nature through fractal geometry is by comparing the fractal dimension of an observed natural phenomenon to a system with a known fractal dimension (whether the known system was created in a laboratory, created mathematically, or is a natural phenomenon that has previously been studied). Similar fractal dimensions are used to support hypotheses about the similarity of the two systems. Thus, the

experimental determination of fractal dimension has become quite important.

Currently, there is no definite theory for determining fractal dimension experimentally. However, what follows is one of the most widely used methods. When real-world data are gathered experimentally, they can usually be plotted as a subset of one of the Euclidean spaces R, R^2, or R^3. Using ε-balls of varying radii, one can "cover" the data with the smallest number of balls possible, obtaining values for ε and $N(A, \varepsilon)$. This procedure is not as simple as it may at first seem, because though we know when a set of points is covered, we can never be sure that the smallest number of balls possible has been used. Nevertheless, with the aid of computers, a very good approximation of $N(A, \varepsilon)$ for a given ε can frequently be obtained.

In R^2, one way of accomplishing this is to superimpose a mesh of grid size ε over the data points, and count the number of squares that contain data points; this procedure has an obvious analogy in R^3. By varying the value of ε, pairs of values for ε and $N(A, \varepsilon)$ can be obtained. Plotting the logs of these values against each other over a range of different values of ε should yield (approximately) a straight line from which the fractal dimension of the set can be determined.

There are some drawbacks to this method. The first problem is that the theory we have developed for fractal dimension applies only to closed, bounded subsets of some Euclidean space. The real world, on the other hand, does not appear to be closed or bounded. Nevertheless, this problem is easily overcome by the fact that once we "measure" the world (i.e., collect a set of data points experimentally) we reduce the real-world phenomenon to a

closed, bounded set (since we can only collect a finite number of pieces of information). Since all of our observations of nature are reduced to a finite set, the closed and bounded assumption fits nicely with our view of the real world.

There is, however, another problem with the experimental method; a problem which is not so easily dismissed. Recall that the definition of fractal dimension invokes the limit as ε approaches zero. But this limiting process cannot be reproduced experimentally. For the fractals created through infinite iterations of function systems, the attractors are known down to the most infinite geometric detail. Moreover, they are completely self-similar at every stage of iteration. However, we do not have this luxury in the real world. We only obtain values of ε and $N(A, \varepsilon)$ for a finite set of nonzero ε values, with no guarantee that the observed behavior will continue as ε approaches 0. For example, consider a piece of coral which appears to be similar in construction to a fractal tree. On the macroscopic level we find numerous self-similar aspects of the geometric shape of the coral. However, as smaller and smaller values of ε are taken, we find that on the cellular level there is no similarity between the shapes of the cells and the shape of the coral. Further analysis reveals that the cell walls themselves even have a different fractal dimension. Moreover, this problem could perpetuate endlessly as we continue to decrease ε.

As another example of this problem, consider the dimension of a ball of string. From a distance, the ball appears to occupy three dimensions in space, so we might conclude its dimension is 3. However, as we move closer to the ball, we see it not as a solid object but as a long string wound around itself, so

we might conclude it has only one dimension. As we get closer still, we see the thickness of the string, and again we conclude it is three-dimensional, but further inspection shows many one-dimensional fibers wound together to make the string. These examples demonstrate the inherent problem with calculating fractal dimension over different ranges of ε, and we have not yet taken ε anywhere near 0.

Since we are plotting ε and $N(A, \varepsilon)$ only over a finite range of ε values, any conclusions drawn about fractal dimension must rely on the gross assumption that the object is infinitely self-similar. Consequently, fractal dimension has been more often used to disprove hypotheses than to conclusively prove them. Nevertheless, if two objects have similar fractal dimensions over the same range of ε values, and if one object is well understood, one may use this information to help explain and analyze the other object.

As an example, consider the dissipation of different pollutants in a laboratory "atmosphere." In an isolated setting, the behavior can be studied in great detail and the fractal dimension of the pollution clouds can be determined over a wide range of ε values. Observing the fractal dimension of real-world pollution clouds over the same values of ε might be a way to determine which pollutant is controlling the clouds' behavior, and which pollutants are escaping into the atmosphere. Similar techniques have been used to liken the growth of certain plants to the formation of crystals in a laboratory, and to extrapolate the behavior of the known system to that of an unknown system.

Two other techniques for experimentally determining fractal dimension are the area-perimeter relationship and the number-resolution relationship. In the area-perimeter relationship, the area of a set of data is computed, and its perimeter is measured several times on a finer and finer scale. In the number-resolution relationship, the number of pixels on a screen is plotted against the resolution of the screen for a variety of resolutions. In both of these techniques, the results (graphs of ordered pairs of perimeters and scale, or numbers and resolutions) are usually compared to those of a known set of data points. If the slopes of these data points (or the logs of the data points) correspond to an object with a known fractal dimension, then the fractal dimension of the unknown object can be estimated. These two techniques are used widely in practice, particularly with photographic images, but again are most commonly used to determine the similarity between two fractal-like objects.

Because the field of fractal geometry is so new, the possiblilties of applications have only begun to be discovered; however, fractal dimension is one aspect that has proven quite useful in many sciences.

J. THE KOCH SNOWFLAKE

An interesting phenomenon that has recently been explained using these ideas is the discrepancy in the recorded lengths of the borders of many countries. In Europe, the reported lengths of the borders between many countries varies greatly, depending on who gives the report. For example, the French records and the German records of the length of the French-German border may differ by several kilometers. The reason for this difference is

simple. A border measured with a ruler from a 1:50,000 scale satellite photograph records much less of the border's detail than a surveying team on the ground records, which again would record much less detail than a government employee on the ground with a ruler. The smaller the device used to measure the border, the more detail will be recorded, and the longer the border will appear to be. This idea is readily illustrated through the analysis of a fractal object called the **Koch snowflake.**

To construct a Koch snowflake, begin with an equilateral triangle, and trisect each of its sides. At each point of trisection, construct another equilateral triangle extending outwards from the original triangle, and trisect each of their external sides (see Figure 3.28). Continuing this geometric construction indefinitely yields the Koch snowflake. While the area inside the Koch snowflake is clearly finite (since the entire snowflake lies inside any circle containing the original triangle) the length of its border is, surprisingly, infinite. To see this, note that at each step the length of the border is increased by a factor of 4/3. Thus, starting with a length of L yields a final length of $\lim_{n\to\infty} L(4/3)^n$, which is an infinite length. If we were to measure the Koch snowflake with rulers of different scales, remarkably different results for the length of the border will be obtained. This is the same phenomenon being experienced measuring border lengths in Europe.

As another real-world application of this concept, many small "resort" lakes entice tourists by advertising "a thousand miles of shoreline," or promises to that effect. Never in such an advertisement will one find a mention of the scale used to arrive at that number, as this is simply an application of the above concepts.

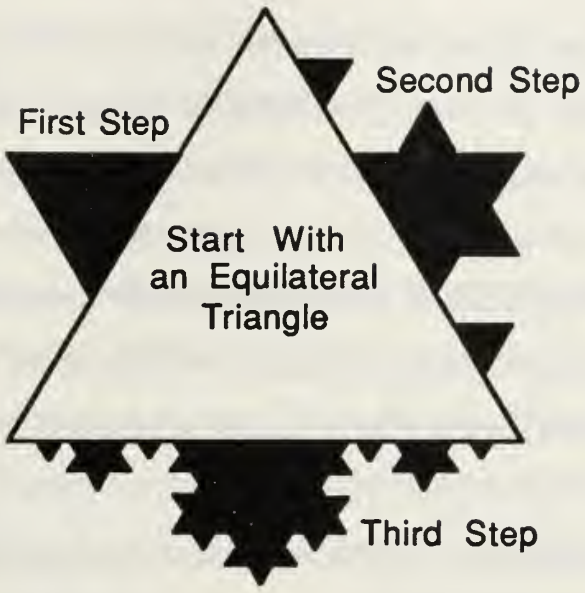

Figure 3.28 Constructing the Koch snowflake.

K. APPLICATIONS OF FRACTAL GEOMETRY

Aside from the applications already mentioned (namely, an aid to computer graphics and the use of fractal dimension to classify and compare similar objects) fractal geometry has so far found few other applications in the physical sciences. As already mentioned, the infinitely-detailed property of fractals does not hold in the real world when objects are analyzed at the molecular and smaller levels. Even Benoit Mandelbrot, the German-born American scientist who gave fractals their name, admits that true fractals do not exist in nature. However, he is also quick to point out that there are no truly straight lines or perfect circles either. While traditional geometries have been inadequate in sufficiently describing all of nature, perhaps fractal geometry, at least on the macroscopic level, may provide yet another approach. When we consider the infinitude of natural objects that have a fractal-like structure, we see that nature is indeed more fractal than it is

Euclidean. From the formation of crystals and snowflakes to the way coral, certain roots, plants, and trees grow; from the shapes of lightning bolts and some electrical discharges to cloud formations, weather patterns, and galactic patterns; from the structure of our lungs and surfaces of our brains, to the patterns of our veins; much of the physical world seems to have a fractal structure to it. A more complete analysis of some of the minutely detailed, self-similar objects in nature may someday be realized through fractal geometry.

One area of mathematics, however, where fractal geometry has unquestionably found a permanent role and distinguished itself is in the study of chaotic dynamical systems. Dynamical (changing) systems are prevalent in the real world and we are finding ever greater numbers of them to be chaotic. In both mathematical models and the physical world, strange sets of data points are being observed that seem to contain self-similarity and infinite detail. Many of these sets of points are, in fact, fractals and with the aid of fractal geometry we can understand them better. Hence, when we realize how much of our universe is modeled by chaotic dynamical systems, we will better appreciate the usefulness of fractal geometry in studying these systems. We take up the study of chaos in the next chapter.

IV. CHAOS

A. INTRODUCTION

We have introduced and studied iterated function systems, from our simplest example where $f^{n+1}(x) = (f^n(x))^2$, to systems that produced very intricate and interesting geometric shapes. The similarity between these systems is that we could predict the exact behavior of the systems after any number of iterations. Unfortunately, in nature this predictability seems to be more the exception than the rule, which has caused scientists frustration for centuries whenever their models of seemingly simple phenomena produced erratic behavior. This observed phenomena in dynamical systems is called *chaos*, a precise definition of which we provide below.

We introduce our study of chaotic dynamical systems by first looking at discrete systems, such as the iterated function systems we have been discussing. There are two reasons for this approach: the first is that discrete systems provide a conceptually simpler setting for understanding the theory, and the second is because many data are collected from the real world in discrete increments. For example, consider a biologist studying the dynamics of a population. Not only will the population change in discrete increments, but the biologist can measure the population only at discrete intervals. In fact, even in continuous real-world behaviors such as beam vibrations, we are able to record measurements only at discrete intervals. Since much of science involves studying a continuous real world based on discrete observations, it

makes sense to begin the study of chaotic dynamical systems with the discrete case.

A further reason to study discrete iterated function systems is that they capture the essence of an important real-world phenomenon known as feedback. Feedback occurs when the present state of a system affects its future state, which in turn means there could be at least a small time delay in the change of the state of a system. For instance, feedback in speaker-microphone systems causes jumps at discrete time intervals because of the time it takes the sound to travel through the system (although we preceive these jumps as continuous when we hear them). Biological systems which are periodic and demonstrate feedback include predator-prey models, the motion of slime molds, light emission by groups of fireflies, glycolysis and photosynthesis, and even the conditions in the brain leading to epileptic seizures. Feedback is also exhibited in electrical circuits, as the defense contractor TRW discovered when chaotic feedback shut down its European computer network. Finally, gaps in the asteroid belts that would be in phase with planetary orbits indicate the effect of feedback in our solar system. These are just a few examples where relatively simple mathematical models can capture important real-world phenomena.

The following five sections are based on Devaney (1989), and provide a background and a setting for studying chaotic dynamical systems. The remainder of the chaper comes from a variety of sources, but in the cases where the references differ on theorems or definitions, the ones used by Devaney are followed. Additionally, many of the examples and applications,

where not specifically referenced, come from Gleick (1988), Moon (1987), Levy (1991), and Briggs (1989).

We begin by developing the mathematics of discrete dynamical systems. Our goal is to understand thoroughly the idea of orbits of points under repeated iteration.

B. GRAPHICAL ANALYSIS OF FIXED POINTS OF MAPS

We have already shown how graphical analysis of a function can reveal information about its fixed points, periodic points, and whether those points are attracting or repelling. We will now introduce the concept of graphical analysis of the nth iterate of a function, denoted by $f^n(x)$. Consider first a very simple function mapping the unit interval to itself $f: [0, 1] \rightarrow [0, 1]$, defined by $f(x) = 2x \pmod 1$. The notation $2x \pmod 1$ means: multiply the input value by two, and consider only the fractional part, i.e., the remainder upon division by one. For example, $f(1.57) = 3.14 \pmod 1 = .14$. This mapping gives $f(x) = 2x$ if $0 \leq x \leq 1/2$ and $f(x) = 2x - 1$ if $1/2 < x \leq 1$. This function is called the **Baker map**, and its graph is shown in Figure 4.1. The Baker map is clearly an onto function, but it is not one-to-one since every y value in $[0, 1]$ is the image of exactly two values of x. Additionally, the fixed points of the Baker map are 0 and 1, both of which are repelling.

Consider the second iterate of the Baker map, $f^2(x) = 4x \pmod 1$. Its behavior can be seen more clearly by the graphical analysis shown in Figure 4.2. Since the interval $[0, 1/2]$ is mapped to $[0, 1]$ under the first iterate, it makes sense that under the second iterate the interval $[0, 1/2]$ will contain a

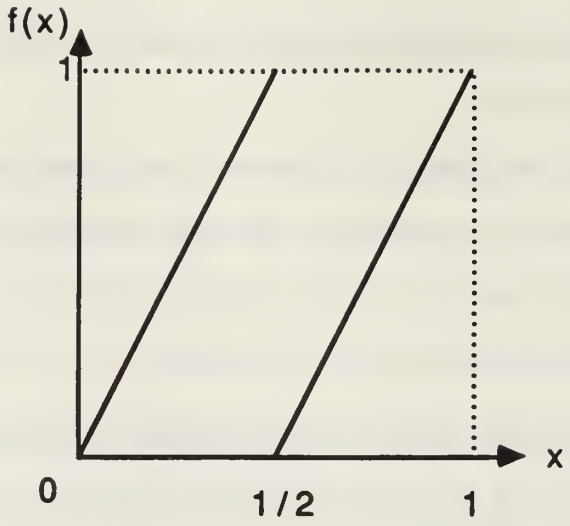

Figure 4.1 The Baker map.

smaller copy of the entire first map. Additionally, the second iterate shows the appearance of two new fixed points, 1/3 and 2/3. This fact demonstrates the existence of a two-cycle in the original map; in fact, it is easy to verify that $f(1/3) = 2/3$ and $f(2/3) = 1/3$. Notice also that the fixed points of $f(x)$, 0 and 1, remain fixed for $f^2(x)$. Observing the fixed points of the nth iterate of a map reveals both the periodic and fixed points of the original map. The third iteration is $f^3(x) = 8x \pmod 1$ and its fixed points, $0, 1/7, 2/7, \ldots, 6/7, 1$ are also shown in Figure 4.2. In the case of the Baker map, we can easily obtain a closed-form expression for its nth iterate: $f^n(x) = 2^n x \pmod 1$. This formula makes the Baker map very easy to graph and analyze, and we will return to it when "chaos" is defined. It may be very difficult, however, to obtain a simple expression for the nth iterate of a map, and it is frequently even more difficult to graph it.

A slightly more complicated map to analyze graphically is the map $f(x) = 4x(1 - x)$, which is a parabola that maps the unit interval to itself. Its

graph and second iterate on the unit interval are shown in Figure 4.4. Again, the appearance of fixed points under $f^2(x)$ shows a two-cycle of $f(x)$, and a simple check shows this two-cycle occurs at the points $x = (5 \pm \sqrt{5})/8$.

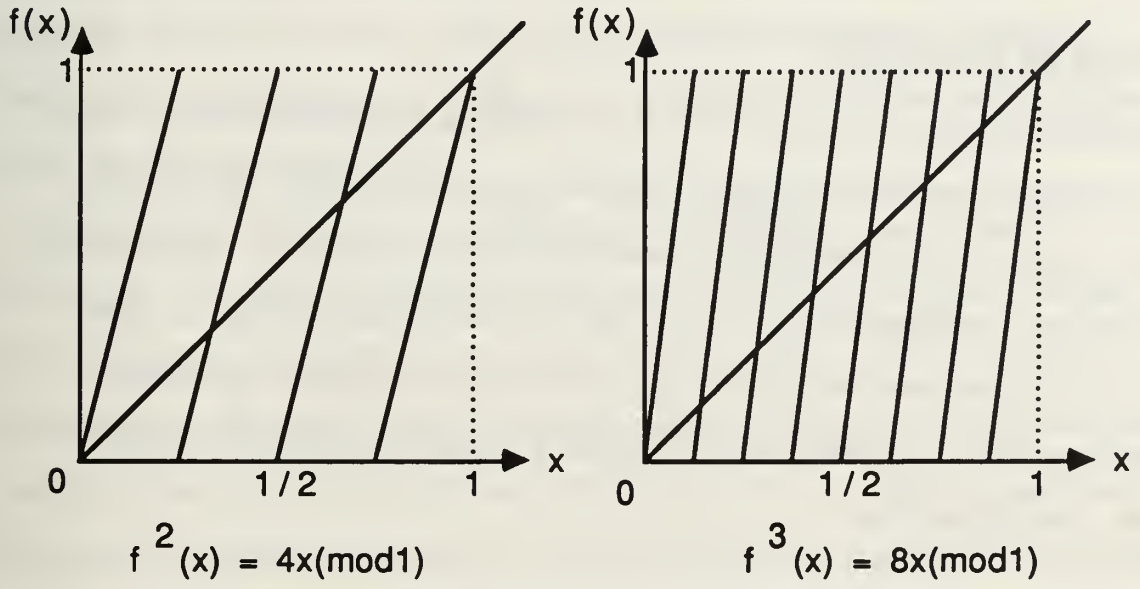

Figure 4.2 The second and third iterates of the Baker map.

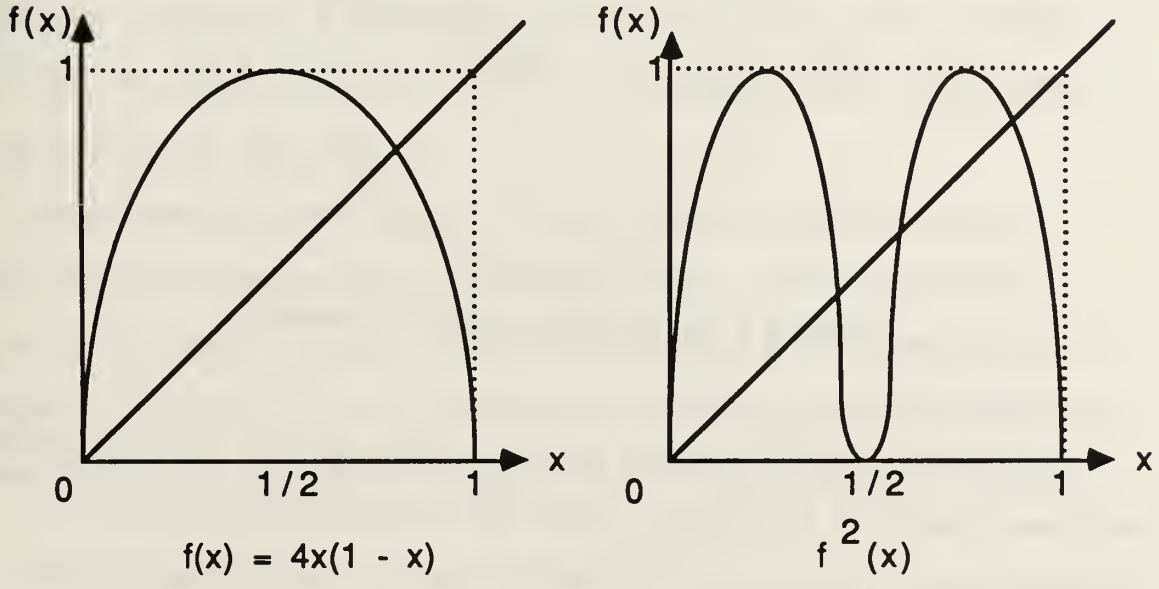

Figure 4.3 The first and second iterates of $f(x) = 4x(1 - x)$.

While this technique is not always easy to use (due to the difficulty of graphing many functions accurately), it can reveal considerable information about the orbits of points in a dynamical system.

C. MAPS OF THE CIRCLE

Consider the map $f: S^1 \to S^1$ with $f(\theta) = 2\theta$ of the unit circle to itself. Here θ is the angle measured in radians (positive counterclockwise) between the positive x-axis and the line joining the origin to any point on S^1. Every point on the circle is represented by the form $\theta \equiv \theta + 2k\pi$ for any integer k. The map f, which doubles the angle of any point on the unit circle, has certain properties which help illustrate and suggest a precise definition of chaotic systems (see Figure 4.4).

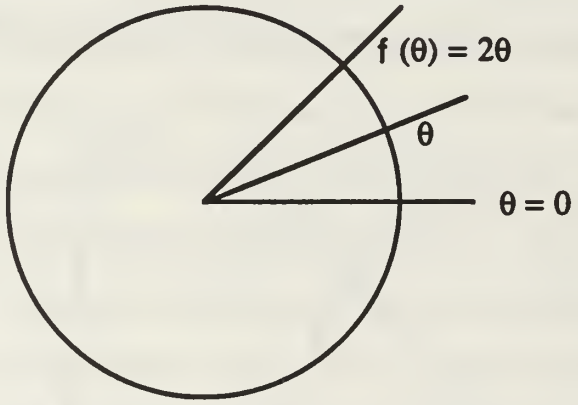

Figure 4.4 The map $f(\theta) = 2\theta$ of the circle S^1.

First notice that $\theta = 0$ is a fixed point of f, since $f(0) = 0$. Additionally, for any point of the form $\theta = 2k\pi/2^n$, $f^n(\theta) = 2k\pi$, so that these points are eventually fixed. The points of the form $\theta = 2k\pi/2^n$ form a dense subset of

S^1, since between any two distinct points on S^1, we can find a point of the form $2k\pi/2^n$. To see this, consider two arbitrary distinct points θ_1 and θ_2 on S^1, where $\theta_1 = 2\pi\alpha_1$ and $\theta_2 = 2\pi\alpha_2$ with $0 \le \alpha_1 < \alpha_2 < 1$. Since $\alpha_2 - \alpha_1 > 0$, there exists an integer m such that $m(\alpha_2 - a_1) > 1$ by the Archimedian property. Let $2^n > m$. Then $2^n(\alpha_2 - \alpha_1) > 1$ implies there exists a positive integer k such that $2^n\alpha_2 > k > 2^n\alpha_1$, or $\theta_2 > 2k\pi/2^n > \theta_1$. Thus, points of the form $2k\pi/2^n$ are dense on S^1.

The periodic points of f are slightly more difficult to find. Since $f^n(\theta) = 2^n\theta$, the point θ is periodic of period n if and only if $2^n\theta = \theta + 2k\pi$ for some integer k. Solving for θ yields $\theta(2^n - 1) = 2k\pi$, so for $0 \le k \le 2^n - 1$, the points $\theta = 2k\pi/(2^n - 1)$ are the periodic points of period n. Using complex analysis, it can be shown that these points are the $(2^n - 1)$ roots of unity, and that they are dense on the circle S^1. Geometrically, this means that the points of period n are those that divide the unit circle into $2^n - 1$ equal segments. This interpretation also shows that they are dense on S^1. This is shown graphically in Figure 4.5 for the points of period two. The point θ_0 trivially has period two since $f^n(\theta_0) = \theta_0$ for all n. The points θ_1 and θ_2 have period two since $f(\theta_1) = \theta_2$ and $f(\theta_2) = \theta_1$.

Another interesting property of f is that for any two points on S^1, no matter how close together they start, iteration under f will separate their iterates by an arbitrary amount. The formal definition of this characteristic is as follows: a mapping $f: S \to S$ has **sensitive dependence on initial conditions** if there exists $\delta > 0$ such that for any point $x \in S$ and any neighborhood N

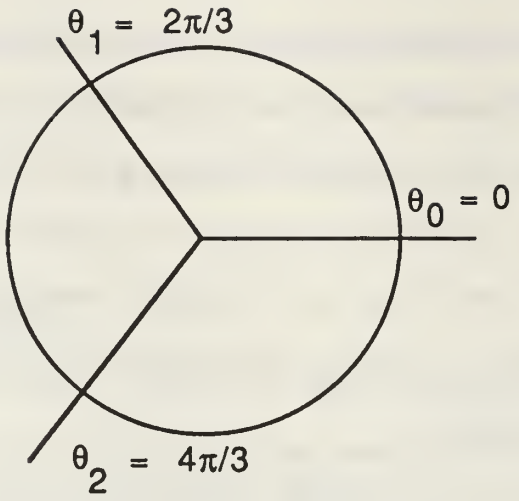

Figure 4.5 Periodic points on S^1 of period two.

of x, there exists a point $y \in N$ and an integer $n \geq 0$ such that $d(f^n(x), f^n(y)) > \delta$. Notice that under this definition, it is not necessary for every point in the neighborhood N to move away from x under repeated iteration. However, in any neighborhood containing x there must exist at least one point that does move away from it. Since our map of the circle $f(\theta) = 2\theta$ doubles the arc-length between any two points, there is always an iteration number n that moves two points at least a given distance δ apart; hence the map $f(\theta) = 2\theta$ has sensitive dependence on initial conditions.

A third interesting property of the map $f(\theta) = 2\theta$ is that for any pair of open arcs $U, V \subset S^1$, there exists a number $n > 0$ such that the intersection of $f^n(U)$ and V is nonempty. The formal definition of this property is as follows: a mapping $f: S \to S$ is called **topologically transitive** if for any pair of open sets $U, V \subset S$ there exists an integer $n > 0$ such that $f^n(U) \cap V \neq \emptyset$. Since any arc U in S^1 is expanded under iteration of $f(\theta) = 2\theta$ until it

covers all of S^1, every arc eventually intersects every other arc in S^1. Thus the mapping $f(\theta) = 2\theta$ is topologically transitive.

The three properties of $f: S^1 \rightarrow S^1$ where $f(\theta) = 2\theta$ define a dynamical system to be chaotic. The formal definition follows.

D. CHAOTIC DYNAMICAL SYSTEMS

A mapping $f: S \rightarrow S$ is called **chaotic** if the following three conditions are met:

1. f has sensitive dependence on initial conditions;
2. f is topologically transitive; and
3. Periodic points are dense in S.

While precise definitions of chaos vary slightly among different disciplines, most of them are at least similar to those we gave here. These three conditions are frequently referred to as **unpredictability, indecomposability,** and **an element of regularity,** respectively. Because of the sensitive dependence on initial conditions, it is almost impossible to predict the orbit of an arbitrary point with any degree of accuracy. This is the unpredictability aspect of chaos, and it makes numerical computations with chaotic dynamical systems virtually impossible (since the slightest rounding error at any step will almost certainly create a point with an entirely different orbit). The indecomposability property comes from the fact that there is no way to decompose the set S into two disjoint subsets that do not eventually interact (since any open subset of S eventually intersects every other open subset of S under iteration of f). Finally, the denseness of the periodic points provides an element of regularity amidst all this chaotic behavior.

Let us now demonstrate that the Baker map is chaotic directly through a graphical analysis. To see that the periodic points are dense, note that the nth iterate $f^n(x) = 2^n x \pmod 1$ has n evenly spaced fixed points in [0, 1] for n arbitrarily large as indicated in Figure 4.2. Figure 4.6 also shows that f is topologically transitive since any open interval (p, q) contains an interval that is mapped to [0, 1]. Finally, Figure 4.7 shows the sensitive dependence on initial conditions, because any two points p and q are moved an arbitrary distance apart under some iteration of f. While this graphical technique is not a proof, it can be supported by a rigorous mathematical argument. It alsodemonstrates the usefulness of analyzing some maps graphically to determine if they are chaotic.

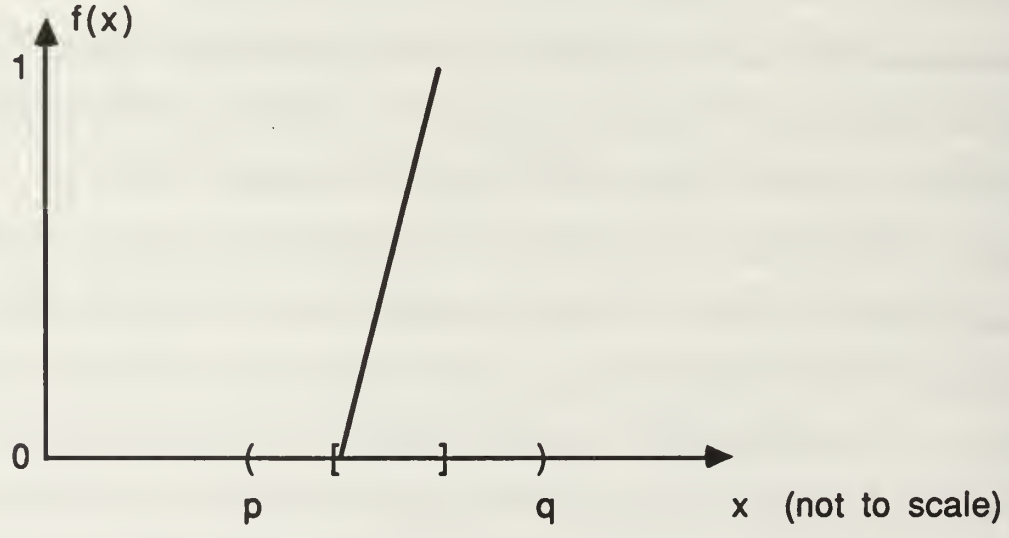

Figure 4.6 Topological transitivity of the Baker map.

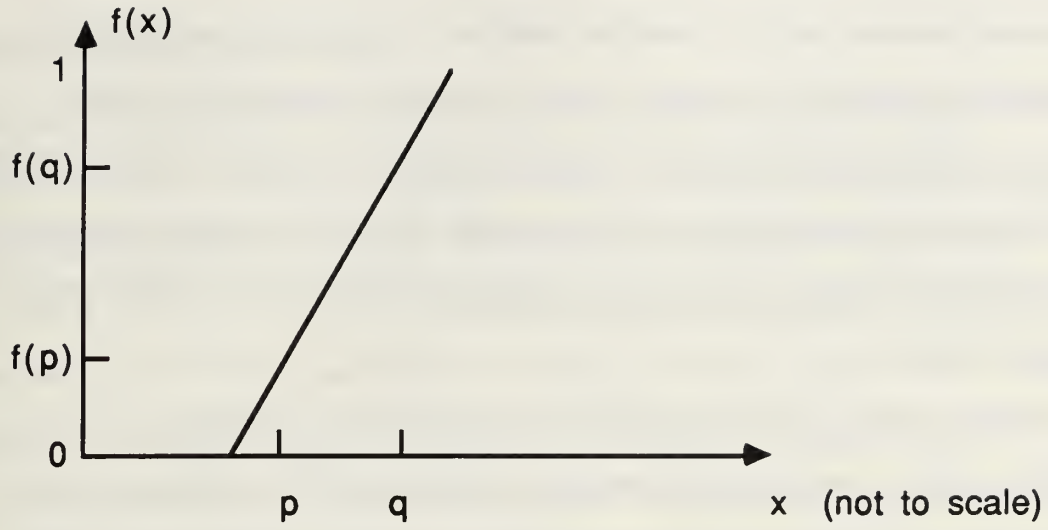

Figure 4.7 Sensitive dependence on initial conditions of the Baker map.

Chaos occurs not only in mathematical equations, but in fact seems to be most prevalent in the real world. While accurately modeling real-world phenomena with mathematical equations is sometimes very difficult, particularly when the physical system is chaotic, it happens that even a simple model crudely representing a physical system may turn out to be chaotic. This happenstance in turn tells us that the more complex physical system is probably also chaotic. We will see this happen when we discuss the Lorenz equations as a crude model for weather prediction. Another example of this chaotic behavior occurs with a simple differential equations model used to describe beam intersections in particle accelerators. The model explains why it is so difficult to predict the action of intersecting beams in particle accelerators (at least using existing models).

There are several examples of physical systems where chaos has been observed, studied, and in some cases, somewhat understood. These systems include turbulence in fluids, thermal convection in gasses, panel flutter on

supersonic aircraft, certain chemical reactions (specifically the Belousof-Zhabotinsky reaction), abnormal cardiac rhythms, nonlinear electrical circuits, biological population dynamics, vibrations of buckled elastic systems (such as beams), geomagnetic field reversals, and even planetary motion. Many of these systems are discussed in Moon (1987), Holden (1986), and Rasband (1990).

Unfortunately, it is very difficult to identify chaotic systems experimentally. One reason is that numerical roundoff in collected data could lead to an erroneous assumption that two data points are the same (indicating a cycle), whereas a slight difference in their actual values may cause their orbits to diverge rapidly (because of sensitive dependence on initial conditions). An even more difficult problem is with distinguishing between a truly chaotic orbit and a cycle with a very long period. To emphasize the significance of this problem we note that the Department of Mechanics at Cornell University requires 4,000 non-cyclic real-world data points before scientists there declare a system as being chaotic. For biologists and economists, who model annual trends, this requirement may seem unreasonable. However, to a helicopter pilot whose life depends on chaotic vibrations not occurring on the main rotor blade, this requirement may not be stringent enough.

While the map $f: S^1 \to S^1$ with $f(\theta) = 2\theta$ provides a simple example of a chaotic map, it by no means represents the extent of chaotic dynamical systems. We turn now to several more interesting chaotic maps.

E. TOPOLOGICAL CONJUGACY

One way to determine if a map is chaotic is to check directly the three conditions of chaos. However, using the concept of topological conjugacy, we can apply the dynamics of a familiar map to those of another function. Two maps, $f: X \to X$ and $g: S \to S$ are said to be **topologically conjugate** if there exists a homeomorphism $h: S \to X$ such that $f(h) = h(g)$. If f and g are topologically conjugate, then $g = h^{-1}fh$, so $g^n = (h^{-1}fh)^n = (h^{-1}fh)(h^{-1}fh) \ldots (h^{-1}fh) = (h^{-1}f)(hh^{-1})f(hh^{-1}) \ldots f(hh^{-1})(fh)$. Since h is a homeomorphism, hh^{-1} is the identity function, hence $g^n = h^{-1}f^nh$, and $hg^n = f^nh$. It follows that f and g share the same dynamical properties; in particular if g is chaotic the same is true of f. This idea is shown in Figure 4.8 with. If $h: S \to X$ is not a homeomorphism (e.g., if it is two-to-one) then f and g are said to be **topologically semi-conjugate**. Nevertheless, if g is chaotic, f is still also chaotic. Since h is two-to-one, the dynamics of f are even more complicated because h introduces even more periodic points or cycles. It suffices that chaotic behavior is preserved through topological semi-conjugacy, which is the extent to which we use this concept.

$$\begin{array}{ccc} S & \xrightarrow{\ g\ } & S \\ \downarrow h & & \downarrow h \\ X & \xrightarrow{\ f\ } & X \end{array}$$

Figure 4.8 Diagram showing topological conjugacy between g and f.

What follows is an example of the usefulness of topological conjugacy. Consider the map $f: R \to R$, where $f(x) = 2x^2 - 1$. We would like to know if

this map is chaotic. We could try to show sensitive dependence on initial conditions, topological transitivity, and search for a dense orbit. Instead we consider the mapping $h(\theta) = \cos\theta$. Since, $\cos 2\theta = 2\cos^2\theta - 1 = f(h(\theta))$, the map f is topologically semi-conjugate (since the cosine function is not one-to-one) with the map $g(\theta) = 2\theta$ when $g: S^1 \to S^1$, as shown in Figure 4.9. Since we know that g is a chaotic map from our previous discussions, and that f shares the same dynamical properties as g, the map $f(x) = 2x^2 - 1$ is also chaotic.

$$\begin{array}{ccc} \theta & \xrightarrow{\ g\ } & 2\theta \\ \downarrow h & & \downarrow h \\ \cos\theta & \xrightarrow{\ f\ } & \cos 2\theta \end{array}$$

Figure 4.9 Topological conjugacy between $g(\theta) = 2\theta$ and $f(x) = 2x^2 - 1$, under $h(\theta) = \cos\theta$.

We frequently use the concept of topological conjugacy to analyze maps suspected to be chaotic. The simplicity of the previous example shows why topological conjugacy is often used.

F. CHAOTIC DYNAMICS ON CODE SPACE

Consider now the shift map $\sigma: \Sigma_2 \to \Sigma_2$ on code space $\Sigma_2 = \{s_1s_2s_3\ldots : s_i \in \{0, 1\}\}$ where $\sigma(s_1s_2s_3\ldots) = s_2s_3s_4\ldots$. We show directly from the definition that this mapping is chaotic.

To see that $\sigma: \Sigma_2 \to \Sigma_2$ has sensitive dependence on initial conditions, consider any point $x = x_1x_2x_3\ldots$ in Σ_2. For a point $y = y_1y_2y_3\ldots$ to be in a

neighborhood of x, it must agree with x for a certain number of digits depending on the size of the neighborhood. Suppose the agreement is with the first n digits; $x_1 = y_1$, $x_2 = y_2, \ldots, x_n = y_n$. If we choose $y_k \neq x_k$ for all $k > n$, then y still lies in the neighborhood of x, but $|\sigma^n(x) - \sigma^n(y)| > \delta$ for all $\delta < 1$. Hence, the point y, which is arbitrarily close to x, moves arbitrarily away from it under iteration of σ. This establishes sensitive dependence on initial conditions.

To show topological transitivity, consider two open sets $U, V \in \Sigma_2$. Since U is open, its points $u_1u_2u_3\ldots u_n\ldots$ in Σ_2 agree for the first n digits, but can differ in any way beyond the nth digit for some positive integer n. Likewise, points in Σ_2 which agree in the first m digits for some m all lie in V. It follows that the point $u_1u_2u_3\ldots u_nv_1v_2v_3\ldots v_m\ldots$ lies in U, while the nth iterate of U under σ contains the point $v_1v_2v_3\ldots v_m\ldots$ in V. Thus the intersection of $\sigma^n(U)$ and V is nonempty for any open sets U and V of Σ_2. Therefore, $\sigma: \Sigma_2 \to \Sigma_2$ is topologically transitive.

To see that the periodic points of σ are dense in Σ_2, we show that for any point $s \in \Sigma_2$ there is a periodic point arbitrarily close to it. Thus, let $s = s_1s_2s_3\ldots$. We want to find a periodic point x that corresponds to s for up to n digits (selecting n arbitrarily large makes the periodic point arbitrarily close to s). Choose the point $x = s_1s_2\ldots s_ns_1s_2\ldots s_n\ldots$, which is a periodic point of period n. The point x is arbitrarily close to S. Hence, the periodic points of σ are dense in Σ_2.

In conclusion, the shift map on code space $\sigma: \Sigma_2 \to \Sigma_2$ is chaotic. Like the map of the unit circle $f(\theta) = 2\theta$, the shift map is quite useful in showing that other maps are chaotic through the use of topological conjugacy. When we

can establish topological conjugacy between the maps f and σ, we will refer to analyzing the behavior of points under f as studying the symbolic dynamics of the map f.

Using the concept of topological conjugacy, it is now an easy exercise to show that the Baker map $f(x) = 2x \pmod 1$ is chaotic. This behavior can be shown through either the shift map on code space or the angle doubling map on the unit circle, which confirms our earlier geometric demonstration of this fact.

G. NEWTON'S METHOD FOR $X^2 = -1$

To see how chaos sometimes arises from mathematical systems, consider the following example from numerical analysis due to Strang (1991). Recall that in finding the zeros of a real valued function $f(x)$ by Newton's method, an initial guess together with the iterative process $x_{n+1} = x_n - f(x_n)/f'(x_n)$ is used. Applying this method to find the roots of $f(x) = x^2 + 1$, the iteration equation reduces to the iterated function system $x_{n+1} = x_n - (x_n^2 + 1)/(2x_n) = (1/2)(x_n - 1/x_n)$. Newton's method converges for most polynomials under certain conditions (e.g., when $f'(x) \neq 0$ and no inflection point occurs between the root and the initial guess). In the case in question, the only roots are $\pm i$ so the system cannot converge on the real line. (But note that $f(x)$ is always strictly positive and has no real roots.)

The behavior of the iterated function system changes dramatically depending on the initial guess. For some values of x_0, this system diverges to infinity; for example, if $x_0 = 1$, then $x_1 = 0$, and $x_2 \to \infty$ (since we cannot divide by zero). On the other hand, if we choose x_0 to be very large, then x_1

is approximately half that value, and the orbit moves towards zero until the $1/x_n$ term makes the next iteration large again. Finally, if we choose $x_0 = 1/\sqrt{3}$, then $x_1 = -1/\sqrt{3}$ and $x_2 = 1/\sqrt{3}$, yielding a two cycle.

Because we are searching for a root value which is nonexistent on the real line, Newton's method exhibits strange behavior for $x^2 + 1 = 0$. We now explore whether this system is chaotic. From trigonometry, $\cot 2\theta = (1/2)(\cot\theta - 1/\cot\theta)$. Using the map $h(\theta) = \cot\theta$ it follows that the map $g(\theta) = 2\theta$ is topologically conjugate to $f(x) = (1/2)(x - 1/x)$, as shown in Figure 4.10.

$$\begin{array}{ccc} \theta & \xrightarrow{\ g\ } & 2\theta \\ \downarrow h & & \downarrow h \\ \cot\theta & \xrightarrow{\ f\ } & \cot 2\theta \end{array}$$

Figure 4.10 Topological conjugacy between $g(\theta) = 2\theta$ and $f(x) = (1/2)(x - 1/x)$ under $h(\theta) = \cot\theta$.

Thus $f(x)$ exhibits the same dynamical properties as $g(\theta) = 2\theta$ on S^1; hence f is chaotic. We can actually observe this behavior graphically by considering the graph of the iterates of f. Letting $x_0 = \cot\theta$, we obtain $x_1 = f(x_0) = \cot 2\theta$, $x_2 = \cot 4\theta$, and in general, $x_n = \cot(2^n\theta)$. Hence, the orbit of x under f is the sequence $\cot\theta, \cot 2\theta, \cot 4\theta \ldots$ for different values of θ corresponding to our initial guess. The graph of $\cot\theta$ is shown in Figure 4.11, and helps to demonstrate the results noted earlier for different initial guesses.

If we start with $x_0 = 1$ we get $\theta_0 = \pi/4$ (since $\cot(\pi/4) = 1$). Then $\theta_1 = \pi/2$, corresponding to $x_1 = 0$ (since $\cot(\pi/2) = 0$). Next $\theta_2 = \pi$ and the orbit for x

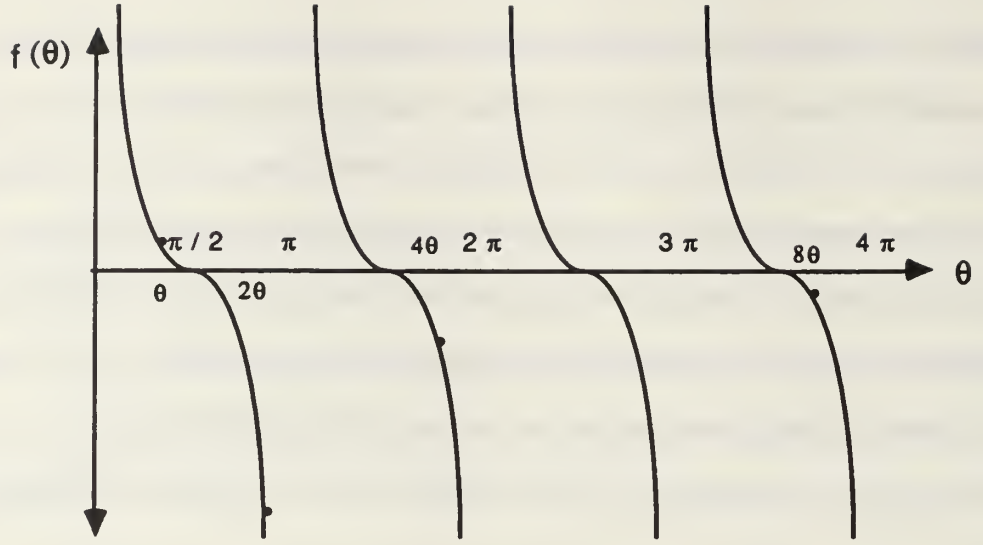

Figure 4.11 The cotangent map and a chaotic orbit.

diverges since $\cot \pi$ is undefined. In fact, any starting value of x_0 corresponding to the form $\theta_0 = k\pi/2^n$, for positive integers k and n, results in an orbit that eventually diverges to ∞. Starting with $x_0 = 1/\sqrt{3}$, we find $\theta_0 = \pi/3$ (since $\cot(\pi/3) = 1/\sqrt{3}$). Then $\theta_1 = 2\theta_0 = 2\pi/3$ gives $x_1 = -1/\sqrt{3}$ (since $\cot(2\pi/3) = -1/\sqrt{3}$). Next $\theta_2 = 2\theta_1 = 4\pi/3 = \theta_0 + \pi$. Since the cotangent function is periodic with period π, $x_2 = 1/\sqrt{3}$. Therefore, this orbit is the two-cycle we observed earlier. In fact, if $\theta_0 = (p/q)\pi$, where p/q is not of the form $k/2^n$, then the orbit eventually cycles. This observation further demonstrates the denseness property of the periodic points. Finally, if θ_0 is an irrational multiple of π, the orbit will never diverge or cycle; in fact, it will be chaotic, as we already knew.

Considering our search for i on the real line, the points that we hope would eventually converge to zero are $y_n = x_n^2 + 1$. If we further analyze these points, we find yet another chaotic map which shares the dynamics of

the one we just studied. For $y_{n+1} = (x_{n+1})^2 + 1 = (1/4)(x_n - 1/x_n)^2 + 1 = (1/4)(x_n^2 + 2 + 1/x_n^2)$. Simplifying algebraically, $y_{n+1} = (1/4)(x_n^2 + 1)^2/x_n^2 = y_n^2/4(y_n - 1)$. If we change variables and let $z = 1/y$, this last equation reduces to $z_{n+1} = 4z_n(1 - z_n)$. The latter iterated function is a member of the quadratic family of maps we study in great detail below, and we know it is chaotic.

As a final illustration of Newton's method, we analyze the system $f(z) = z^4 - 1$ in the complex plane. Since this equation has four roots, ± 1 and $\pm i$, depending on the initial value, the iterations could end up at any one of the four roots. The results encountered with computers when searching for these roots has not always been predictable, depending on the initial value. Now each of these roots has around it a "basin of attraction" inside of which all initial values converge to that root upon repeated iteration. However, along the diagonals $Re(z) = Im(z)$ and $Re(z) = -Im(z)$, there are regions that give rise to chaotic behavior. To see this, color the complex plane based on which initial values converge to which root: that is color all the points that converge to 1 blue, to -1 green, to i yellow, and to -i red. There are four large basins of attraction with infinitely detailed, multicolored borders between them. In fact, there is no clearly defined boundary between any two basins of attraction: between any blue and red region, there is a yellow and a green one. A schematic of the complex plane showing the basins of attraction and the chaotic regions is shown in Figure 4.12. A detailed color image of this figure can be found in Gleick (1988, p. 114).

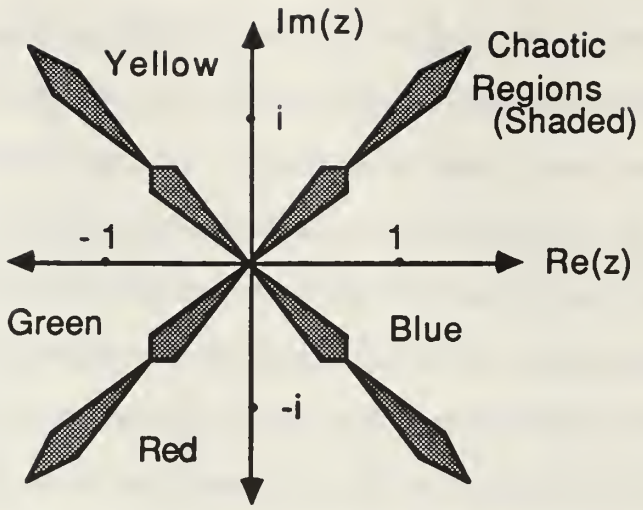

Figure 4.12 Schematic of the complex plane indicating the behavior under Newton's method for $z^4 + 1$ and the (colored) basins of attraction.

H. THE QUADRATIC FAMILY OF MAPS

In this section we study a specific case of maps from the quadratic family $f(x) = ax^2 + bx + c$. We discuss a model from population biology to develop this map, which comes from Briggs (1989). The mathematics in this section is from the book by Devaney (1989) and the article by Devaney (1989), while some of the more detailed results, particularly about the Feigenbaum constant, are from Rasband (1990).

Consider a population of moths which live for one season, lay their eggs, and then die, all in a fixed and limited environment. If we do not consider the interaction of the moths with the environment or other species, we can model the dynamics of the moth population by assuming that the size of the population increases after each life cycle in constant proportion to the size of the population during the previous cycle. If we let the constant of

proportionality equal B (for birthrate), then for each population cycle we have $x_{n+1} = Bx_n$, where x represents the population and n represents the time period. Given only these conditions together with $B > 1$, this model predicts the population increases forever and eventually becomes arbitrarily large. Clearly no population in the real world is represented by this model, since no limited environment can sustain an infinite population. Thus it is necessary to adjust the model.

Suppose then that as a population increases above the maximum sustainable capacity of the environment, the competition for resources causes some of the population to die off or be killed. Hence a particularly large population might cause the population to actually decrease in the next cycle. If we scale our model so that $0 < x < 1$, with 1 being the capacity of the environment, then incorporating the multiplicative factor $(1 - x)$ into our model could account for the limitations of the environment on large populations. We now adjust B to account for the birthrate (and deathrate) in a fixed environment and denote its new value by λ, obtaining the refined model $x_{n+1} = \lambda x_n(1 - x_n)$. This model shows that for small populations, the population growth is close to the original model $x_{n+1} = Bx_n$. However, as the population approaches the maximum sustainable capacity of the environment, its growth starts to decline because of the $(1 - x_n)$ factor.

The equation $x_{n+1} = \lambda x_n(1 - x_n)$, $\lambda > 0$, is called the **logistic equation**. It is perhaps one of the simplest population models to represent the interaction between a species and its fixed environment (although it excludes interactions with other species). Allowing x to vary between zero and one scales the model to any population size. Restricting the parameter $0 \leq \lambda \leq 4$,

we see below that the model never allows the population to increase above the maximum capacity of the environment. The logistic equation has been used by biologists for many years, and it has been a major goal to understand populations that are modelled by this equation. One methodology has been to determine experimentally a value of λ based on the species and the environment, and then to observe the population over time to see if it behaves according to the model. This simple model is a fairly good predictor for a simple species like bacteria or yeast growing in a culture, but it has not always correctly predicted the observed results for more highly developed species, like mammals. Some populations have been observed to settle down to a fixed value for all time increments, others have been seen to cycle between two, four, or even larger periods, and still others have demonstrated completely unpredictable behavior that fit no known model at all. What happens is that, for certain values of λ, the logistic equation $f(x_n) = \lambda x_n(1 - x_n)$ becomes chaotic. Thus it is impossible to predict its behavior even after a few iterations. We next examine this map mathematically.

First note that for values of x_0 equal to 0 or 1, the next iteration $x_1 = 0$ is a fixed point of this map, since $f(0) = 0$. Also, for values of x_0 outside the interval $[0, 1]$, x_1 becomes negative (which does not make physical sense since we are studying populations). Furthermore, all negative values diverge to $-\infty$ under repeated iteration of f. Hence, for practical reasons we concentrate on the dynamics of this system on the interval $[0, 1]$.

Suppose $0 < \lambda < 1$. Then $x = 0$ is the only fixed point of the system (see Figure 4.13). Moreover, all values of $x_0 \in [0, 1]$ tend to 0 under iteration of f. (Recall that if $|f'(x)| < 1$ for a fixed point x, then the fixed point is

attracting. Since $f'(x) = \lambda(1 - 2x)$, for $\lambda < 1$, we obtain $f'(0) = \lambda < 1$; so 0 is an attracting fixed point). The physical interpretation of this result is that the population does not reproduce fast enough to sustain itself and eventually dies out.

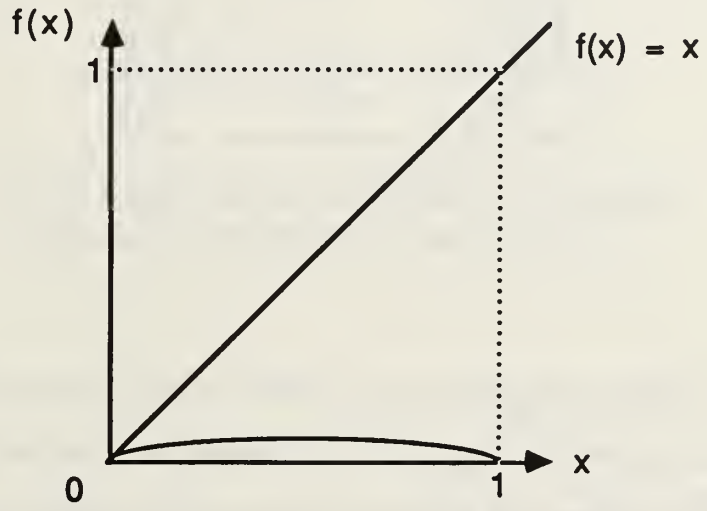

Figure 4.13 The logistic equation for $\lambda < 1$.

Now increase λ so that it is greater than 1. Then 0 becomes a repelling fixed point since $f'(0) = \lambda > 1$. However, note also the introduction of another fixed point p in the interval [0, 1], given by $p = 1 - 1/\lambda$ (see Figure 4.14). If we consider values of λ such that $1 < \lambda < 3$, we see that the point p is an attracting fixed point since $|f'(p)| = |\lambda(1 - 2 + 2/\lambda)| = |2 - \lambda| < 1$. Hence, for values of λ between 1 and 3, 0 is a repelling fixed point and $p = 1 - 1/\lambda$ is an attracting fixed point. This means that our population will always settle down to a fixed value and not vary in cycles.

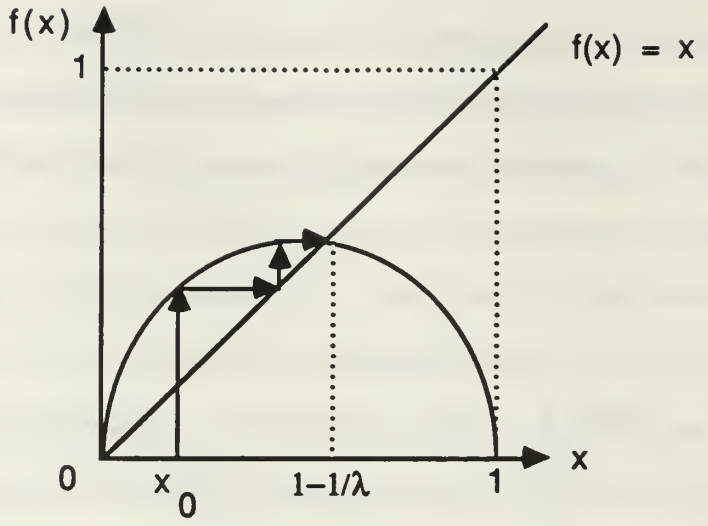

Figure 4.14 The logistic equation for $1 < \lambda < 3$.

Now let λ jump to the value $\lambda = 4$. Since $x_{n+1} = 4x_n(1 - x_n)$ we know that this map is chaotic because of our discussion of Newton's method for $z^2 = -1$. We now demonstrate this idea graphically to clarify further what is actually happening. Consider the graph of $f(x) = 4x(1 - x)$ on the interval $[0, 1]$ in Figure 4.15. The graph represents a two-to-one and onto function. Hence, the second iterate $f^2(x)$ creates a condensed copy of $f(x)$ on each interval $[0, 1/2]$ and $[1/2, 1]$. Moreover, that $f^2(x)$ has four fixed points (see Figure 4.16). The third iteration $f^3(x)$ is also shown in Figure 4.16. Observe that we are creating a graphical situation similar to the Baker map studied earlier (with parabolic arcs rather than straight lines). The graphs reveal that $f(x)$ has sensitive dependence on initial conditions, topological transitivity, and a dense orbit; so it is, in fact, chaotic. This analysis explains why, for some populations, the observed values fluctuate wildly, despite a seemingly simple mathematical model describing the behavior.

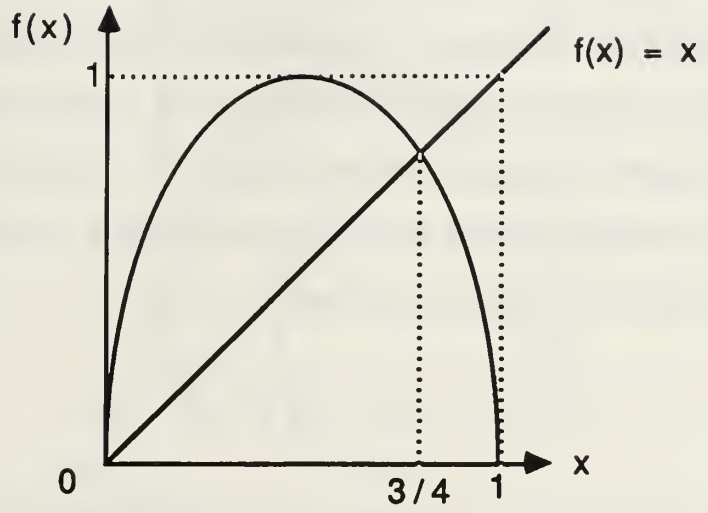

Figure 4.15 The logistic equation for $\lambda = 4$.

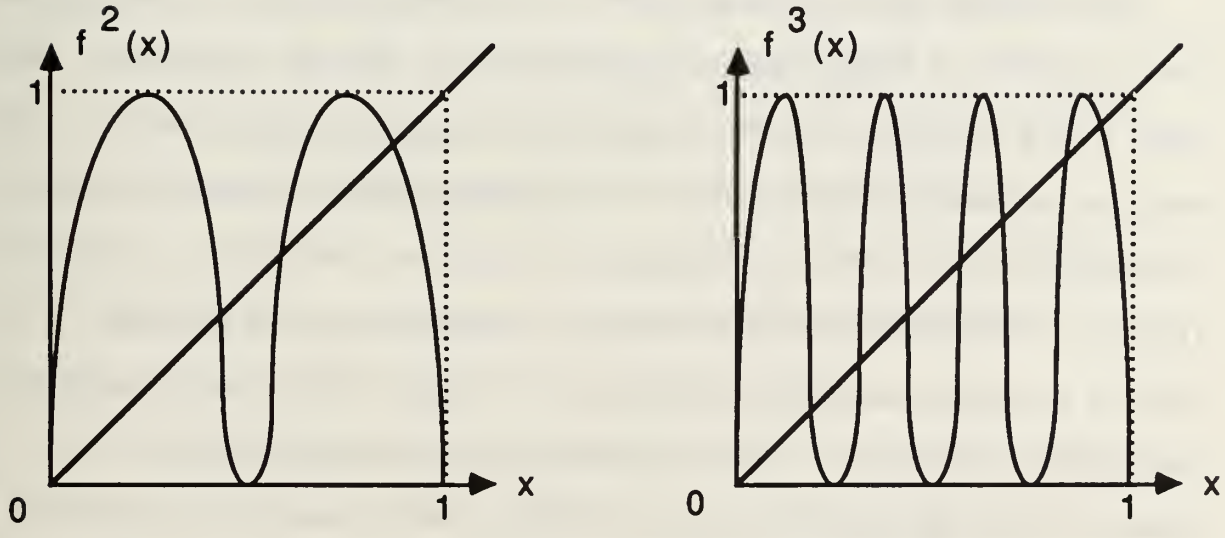

Figure 4.16 The second and third iterates of the logistic equation for $\lambda = 4$.

Another way to show that the map $f(x) = 4x(1 - x)$ is chaotic is to use the concept of topological conjugacy. Recalling the trigonometric identity $\sin^2 2\theta = 4\sin^2\theta(1 - \sin^2\theta)$, the mapping $h(\theta) = \sin^2\theta$ shows that $f(x) = 4x(1 - x)$ is topologically semi-conjugate with the map of the circle $g(\theta) = 2\theta$. This commutative diagram is shown in Figure 4.17. This is yet another way to show that the map $f(x) = 4x(1 - x)$ is chaotic.

$$\begin{array}{ccc} \theta & \xrightarrow{\ g\ } & 2\theta \\ \downarrow h & & \downarrow h \\ \sin^2\theta & \xrightarrow{\ f\ } & \sin^2 2\theta \end{array}$$

Figure 4.17 Topological conjugacy between $g(\theta) = 2\theta$ and $f(x) = 4x(1 - x)$.

Now examine what happens when $3 < \lambda < 4$. We observed an attracting fixed point for $\lambda < 3$, and chaotic behavior for $\lambda = 4$. Thus as λ increases from 3 to 4 the dynamics of the system must change dramatically to produce this chaotic behavior. For $\lambda = 3$, the magnitude of the derivative at the fixed point $p = 1 - 1/\lambda = 2/3$ is equal to $|f'(2/3)| = |2 - \lambda| = 1$; for $\lambda > 3$, $|f'(p)| > 1$. This means that as the value of λ passes through 3, the fixed point p goes from attracting to repelling. This result implies that both of the fixed points, 0 and $p = 1 - 1/\lambda$ are repelling. With both fixed points repelling, one might think the orbit will fluctuate wildly within the interval [0, 1]. But what really occurs is the creation of an attractive two-cycle, as shown in Figure 4.18. This splitting of a fixed point into an attractive cycle is called a *period-doubling bifurcation*.

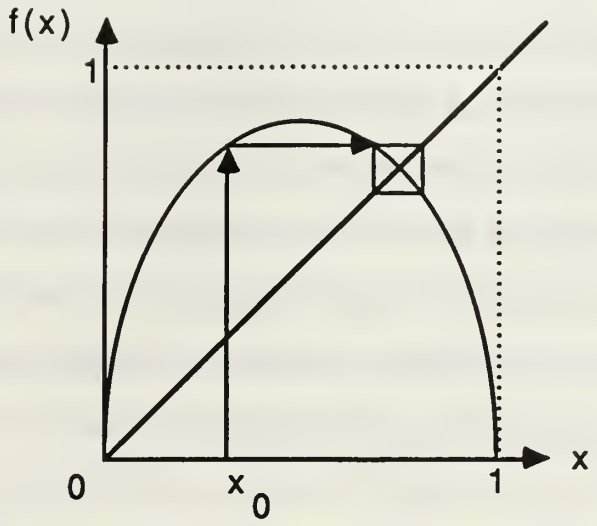

Figure 4.18 An attractive two-cycle in the logistic equation for $\lambda > 3$.

The two-cycle previously described is one of the phenomena that puzzled biologists for many years. Physically, a small population would flourish and breed very rapidly, since the environment could easily support it. After a time a large population would be created in the next cycle. This population would breed more sluggishly because it was nearer the capacity of the environment. Furthermore, the population would not approach its fixed value (since that value was repelling). Even when biologists altered the population it would still tend back to the attractive cycle.

As we continue to increase the value of λ beyond $\lambda = 3$, we find that at $\lambda \approx 3.4495$ the two-cycle becomes repelling and an attractive four-cycle is created. Now, as with $\lambda = 3$, the derivative $|(f^n)'(p)|$ for the cyclic points p_1 and p_2 is passing through 1, and the two-cycle becomes repelling; another bifurcation of the system has occured.

If we continue to increase λ, the next bifurcation occurs at $\lambda \approx 3.544$, where the four-cycle becomes repelling and an attractive eight-cycle is created.

Plotting the values of the parameter λ against the values of the attractor (the attracting periodic points of period two, four, eight, . . .) shows how each cycle bifurcates into another cycle with twice the period as the previous one. Such a **bifurcation diagram** is shown in Figure 4.19 for $2.9 < \lambda < 3.9$. The parameter values at which the next four bifurcations occur are $\lambda \approx 3.5644, 3.5688, 3.5697,$ and 3.5699. As one can see, these numbers are getting closer together and they converge to a value where the map becomes chaotic.

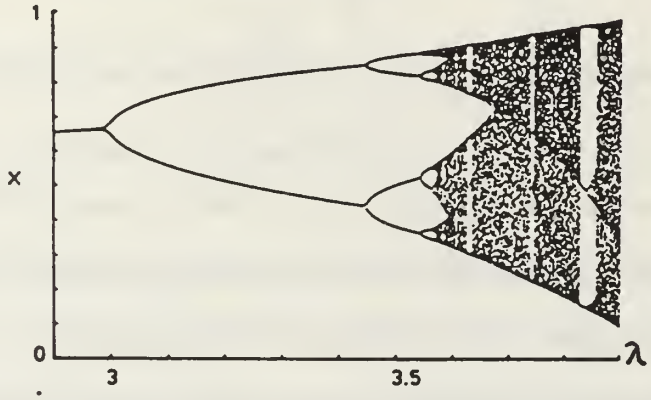

Figure 4.19 The bifurcation diagram for the logistic equation as λ increases. The figure is from Holden (1986, p. 46).

The apparent convergence of these numbers led to the discovery of a constant that appears to be almost universal in dynamical systems. Denote the values of λ at which bifurcations occur by λ_k for $k = 1, 2, 3, \ldots$, denote the value to which they converge by λ_∞, and consider the following. Figure 4.20 shows the second iterate of the logistic equation. Note that it contains a scaled copy of the first iterate, and that each further iterate will again contain a scaled copy of the original map. This geometric observation provides insight to the phenomenon of period-doubling. At $\lambda = \lambda_{k-1}$, a period-doubling occurs in the system that lasts for a duration (in λ) of $(\lambda_k - \lambda_{k-1})$. The next

period doubling occurs at $\lambda = \lambda_k$, and lasts for a λ-duration of $(\lambda_{k+1} - \lambda_k)$. It is not unreasonable to expect a "scaling" in the successive λ-lapses based on the graphical observation. Hence, if we assume the convergence of $(\lambda_\infty - \lambda_k)$ to be geometric, then $\lambda_\infty - \lambda_k = c/\delta^k$ where c is a constant and δ is a constant greater than one. Using this to solve for δ in terms of the λ_k, we see that $\delta = (\lambda_k - \lambda_{k-1})/(\lambda_{k+1} - \lambda_k)$. From our previously computed values of λ_k, we see that $\delta \approx 4.6692$, which is called the **Feigenbaum constant**, named after the American scientist and mathematician Mitchell Feigenbaum, who discovered it. Using this constant to solve for λ_∞, we find that $\lambda_\infty \approx 3.5699456$. Hence, with $\lambda = \lambda_\infty$, the logistic map is chaotic. (As an aside, the Feigenbaum constant has been found in many different dynamical systems. For example, models for electrical circuitry, optical systems, and economic cycles, as well as physical systems such as chemical reactions, erratic heart behavior, and even dripping faucets, have all exhibited the Feigenbaum constant when viewed in the proper phase space. With its value now known, the number of dynamical systems where the Feigenbaum constant is found is increasing quite rapidly.)

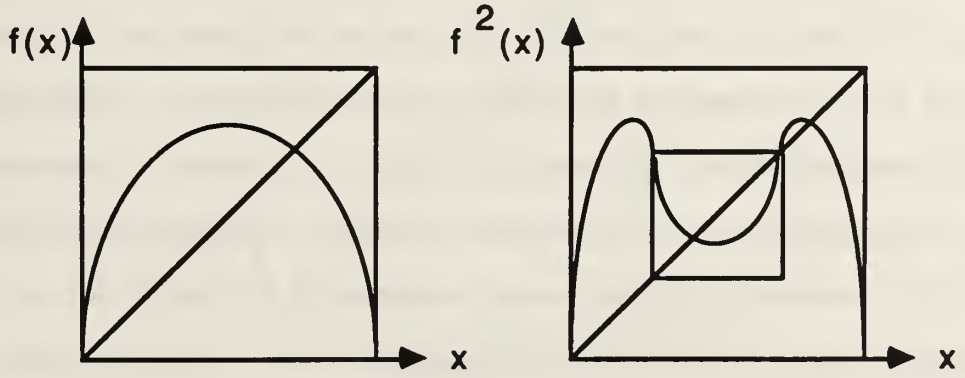

Figure 4.20 The first and second iterates of the logistic equation.

This "period-doubling route to chaos" was one of the first routes to chaos understood by mathematicians. It has helped biologists understand the many different results associated with systems modeled with the logistic map. While this discussion about bifurcations and the period-doubling route to chaos has been mostly geometric and intuitive, there is a rigorous underlying mathematical theory validating these concepts. However, this theory requires an understanding of "kneading theory," an elaborate version of symbolic dynamics that is beyond the scope of this thesis.

While the logistic map is very restrictive (in that it models only a single species interacting with a fixed environment) it still provides a great deal of information about population dynamics. When the model is further refined with the introduction of another species (i.e., a predator-prey or plant-herbivore model), the possibility for chaotic behavior increases with the complexity of the model. Also, the logistic equation has been used by medical scientists to model the spread of infectious diseases (the x term being the number who are contagious and the $(1 - x)$ term being the number who have developed immunity), and even by sociologists to model the spread of rumors.

We now continue our mathematical analysis of the logistic map beyond the ranges of λ that correspond to physical population systems. Specifically, we examine what occurs with the map $x_{n+1} = \lambda x_n(1 - x_n)$ when λ increases above 4. The simplest way to observe the dynamics is graphically, as shown in Figure 4.21. As before, all initial values outside of [0, 1] diverge to $-\infty$ under infinite iteration. Notice now, however, that there is a set of points inside [0, 1] mapped outside the interval after the first iteration; thus they

diverge to $-\infty$. Examining the second iteration, we see that the two subintervals remaining after the first iteration are mapped to [0, 1]; each will then contain a smaller copy of the original map, and hence a subinterval mapped outside of [0, 1] (see Figure 4.22). In fact, if we consider the set of points which do not eventually diverge to $-\infty$, by this construction we obtain a Cantor set of points remaining in the interval after infinite iteration for $\lambda > 4$. Values of the interval [0, 1] mapped outside of the unit interval, and hence diverging to $-\infty$, are said to be values that **escape** under iteration of f.

The value of λ for which the interval (1/3, 2/3) escapes under the first iteration of f can be found by letting $1 = \lambda(1/3)(1 - 1/3)$, which yields $\lambda = 9/2$. However, iterating the logistic equation with $\lambda = 9/2$ will not yield the classical Cantor set, since the subintervals created in the second iteration will be skewed and will not be exact middle-thirds intervals. In fact, the point

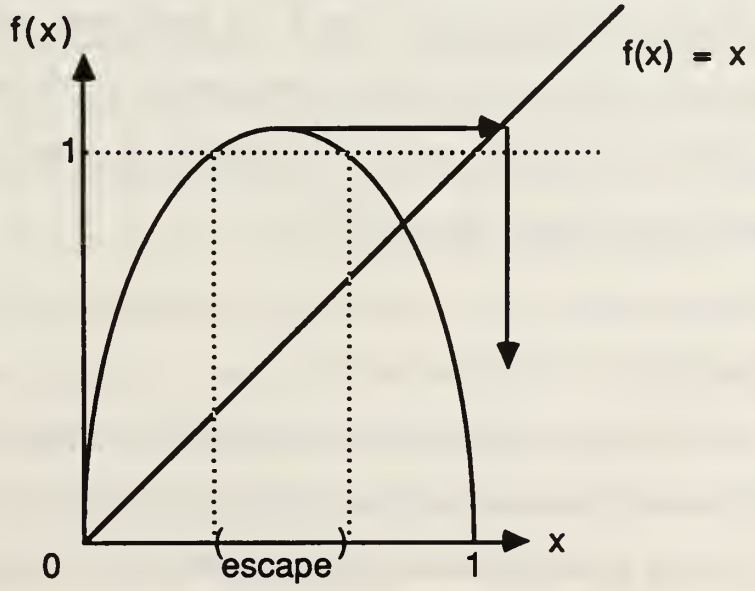

Figure 4.21 The logistic equation with $\lambda > 4$ and the interval that escapes.

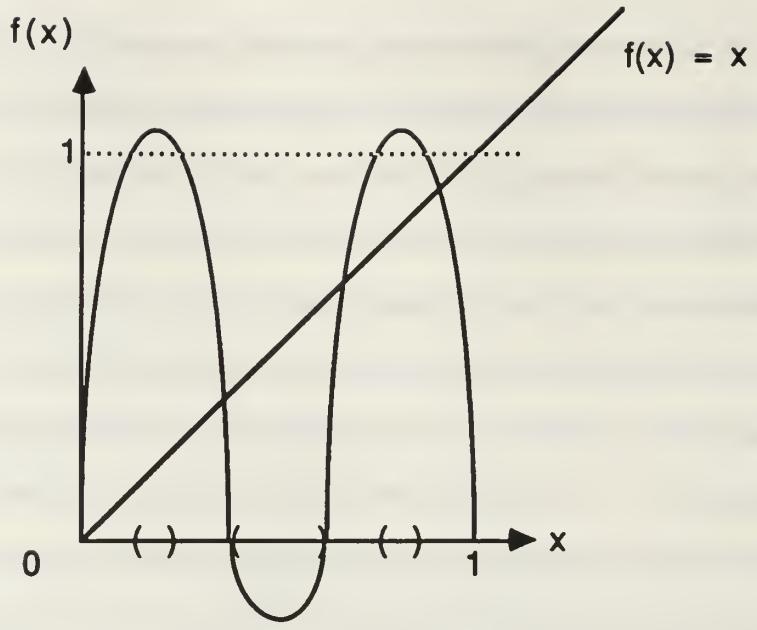

Figure 4.22 The second iterate of the logistic equation with $\lambda > 4$.

$x = 1/9$ escapes under the second iteration. However, we can also see how a Cantor-like set of points remains in the set. Considering $\lambda = 9/2$, we see that $f(1/3) = 1$, and since $f(1) = 0$, with 0 being a fixed point of the map, the point $1/3$ forever remains fixed in the set. For $\lambda > 4$, the logistic equation is chaotic on a Cantor subset of [0, 1], while on the complement of this set, all orbits diverge predictably. This contrasts with the logistic equation with $l = 4$, which is chaotic on the entire interval [0, 1].

I. BIFURCATIONS

In the last section we introduced the concept of bifurcations as period-doubling phenomenon leading to chaos in the logistic map. We now study the general concept of bifurcation and its interpretation in the real world. The mathematical theory in this chapter is from Rasband (1990),

Guckenheimer (1990), and Seydel (1988), while the particular definitions and theorems are from Devaney (1989). The examples are taken from Moon (1987) and Holden (1986).

Bifurcations are normally associated with a change in the physical state of a system. In population dynamics, one bifurcation is exhibited by a population changing from tending towards an attracting fixed point to tending towards an attractive cycle. Changes in the states of other physical systems (such as water cooling until it freezes, a beam bending under a load until it buckles, a balloon being blown up until it bursts, or even the crash of the stock market) are also examples of bifurcations. Clearly, some of these behaviors have such complex and intricate mathematical models that it would be virtually impossible to quantify them and determine the exact nature of the bifurcation. Nevertheless, there are several systems which have been modeled successfully and for which the mathematical analysis of bifurcation accurately predicts gross changes in the state of the system.

Mathematically, there are several types of bifurcations. First we discuss those that occur in one dimension (as with our previous example of the logistic equation). In the logistic equation, the bifurcation occurs as the parameter value λ is varied until it passes through certain values which changed dramatically the behavior of the system. This example suggests we consider families of functions of real variables, where the functions depend "smoothly" on a parameter. Thus, define $g(x, \lambda) = f_\lambda(x)$, where $f(x)$ is a C^∞ function of x for each fixed λ, and $g(x, \lambda)$ depends smoothly on λ. The logistic equation $g(x, \lambda) = f_\lambda(x) = \lambda x(1 - x)$ is one such family.

We first present a general result about bifurcation theory that applies to all one-dimensional bifurcations. Then we give specific examples of typical bifurcations and physical systems that behave similarly in the real world. The general result is that bifurcations occur only near non-hyperbolic fixed and periodic points. (Recall that a hyperbolic periodic point is one where $|(f^n)'(p)| \neq 1$, where n is the prime period of the point; for a fixed point, $n = 1$.) While the theory applies to both fixed and periodic points, we present it here only for fixed points. Recalling further that a fixed point p is an attractor if $|f'(p)| < 1$, and a repellor if $|f'(p)| > 1$, we would expect to find interesting behavior as $|f'(p)|$ passes through 1. Such points p are the only points where bifurcations can and do occur.

BIFURCATION THEOREM. Let f_λ be a one parameter family of functions, with p a hyperbolic fixed point for some fixed λ_0, i.e., $f_{\lambda_0}(p) = p$ and $f'_{\lambda_0}(p) \neq 1$. Then there exist intervals, I about p and N about λ_0, and a smooth function $h: N \to I$ such that $h(\lambda_0) = p$ and $f_\lambda(h(\lambda)) = h(\lambda)$. Moreover, f_λ has no other fixed points in I.

PROOF. Consider the function defined by $G(x, \lambda) = f_\lambda(x) - x$. By hypothesis, $G(p, \lambda_0) = 0$, and $\partial G/\partial x|_{(p, \lambda_0)} = f_{\lambda_0}'(p) - 1 \neq 0$. By the Implicit Function Theorem, there are intervals, I about p and N about λ_0, and a smooth function $h: N \to I$ such that $h(\lambda_0) = p$, and $G(h(\lambda), \lambda) = 0$ for all $\lambda \in N$. Moreover, $g(x, \lambda) \neq 0$ unless $x = h(\lambda)$. This completes the proof.

A graphical demonstration of this theorem is shown in Figure 4.23. The graph shows that nearby graphs $f_{\lambda 1}$ and $f_{\lambda 2}$ must have the same property for a sufficiently small interval about λ_0, because the graph of f_λ meets the line $y = x$ at an angle at (p, p) and since f varies smoothly with λ. Hence, there is only one fixed point near p for λ in some neighborhood of λ_0. At this point there is a function g_λ that is topologically conjugate to f_λ via the map $H_\lambda(x) = x - h(\lambda)$ for which the origin is always a fixed point. This fact allows us to study maps with 0 as a fixed point, and apply the results to any map f_λ with a nonzero fixed point. (To see this, consider f_λ with $f(h(\lambda)) = f(\lambda)$. If $g_\lambda(x) = f_\lambda(x + h(\lambda)) - h(\lambda)$, then $g_\lambda(0) = f_\lambda(h(\lambda)) - h(\lambda) = 0$ for all λ, so 0 is always a fixed point for g.) Thus it suffices to present our complete results for one-dimensional bifurcations using functions whose fixed point is the origin.

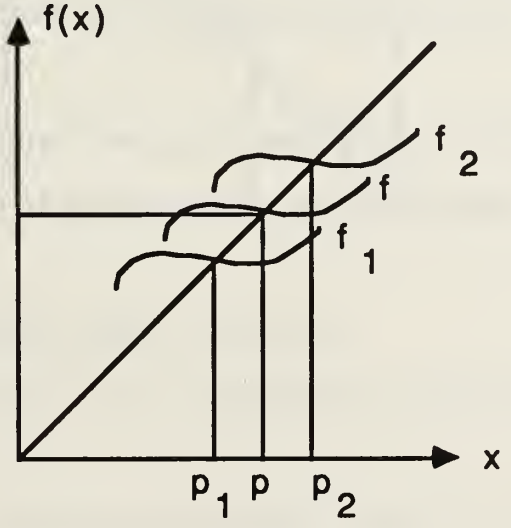

Figure 4.23 Schematic of the Bifurcation theorem.

Now that we know that bifurcations occur at non-hyperbolic periodic points, we are prepared to study specific types of one dimensional bifurcations. The first, and perhaps mathematically the simplest type, is the tangent bifurcation (also called the **saddle-node**). The family $f_\lambda(x) = \lambda e^x$, for $\lambda > 0$, has a bifurcation at $x = 0$ when $\lambda = 1/e$. If $\lambda > 1/e$ the function f_λ has no fixed points, and $f_\lambda^n(x) \to \infty$ for all x. When $\lambda = 1/e$, the function has one fixed point at $p = 1$. This point is attracting from the left and repelling from the right. This last condition, which may occur for a non-hyperbolic fixed point, is called **semi-stable**. For values of $\lambda < 1/e$, f_λ has two fixed points, p_1 and p_2, with p_1 attracting and p_2 repelling. Graphs of these three cases are displayed in Figure 4.24, and the graphical analyses are left to the reader.

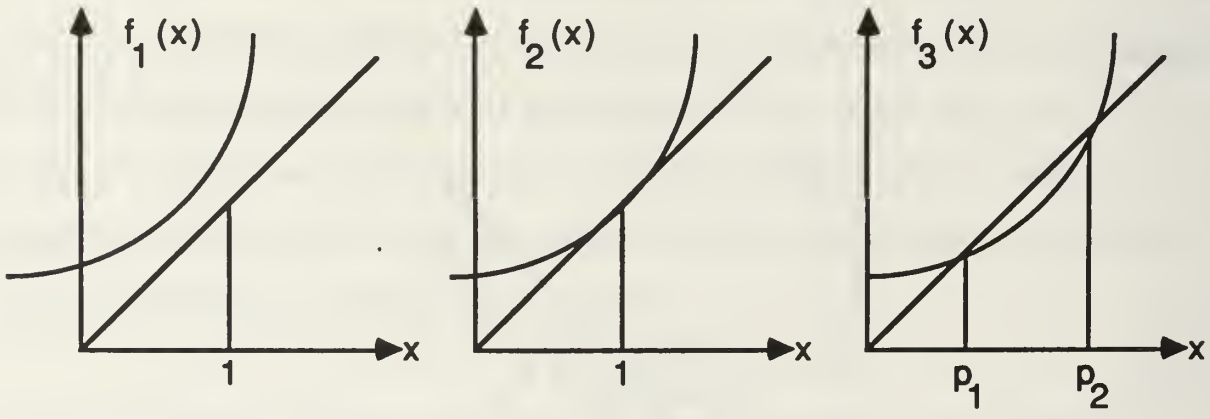

Figure 4.24 Graphs of $f(x) = \lambda e^x$ with (1) $\lambda < 1/e$ (2) $\lambda = 1/e$ (3) $\lambda > 1/e$.

We can see that this bifurcation occurs as the graph of f_λ becomes tangent to the line $y = x$. First one, then two, fixed points emerge as the parameter passes through the critical value. We can also verify that this

bifurcation occurrs at a non-hyperbolic fixed point $p = 1$, since $f_{1/e}'(1) = (1/e)e^1 = 1$. This bifurcation gets its name from the way f_λ approaches the line $y = x$ tangentially. The bifurcation diagram for the function $f_{1/e}$, plotting the fixed points p on the vertical axis versus λ is shown in Figure 4.25.

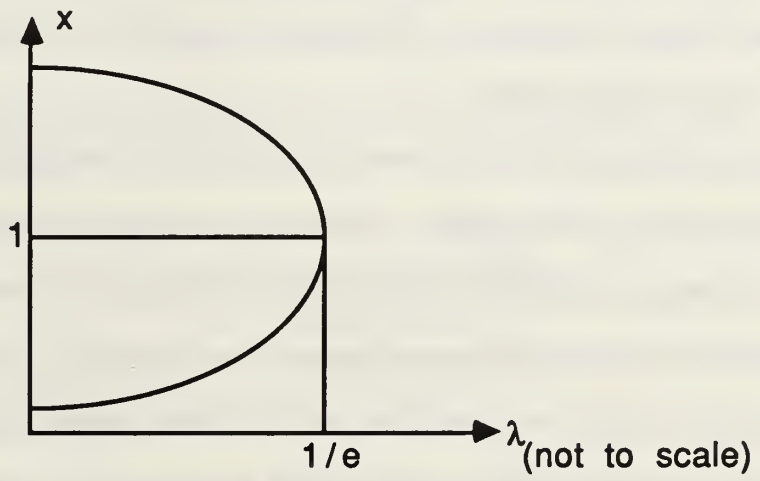

Figure 4.25 Bifurcation diagram for $f(x) = \lambda e^x$.

The mathematical conditions guaranteeing the occurrence of a tangent bifurcation, and the resulting bifurcation, are as follows.

THE TANGENT BIFURCATION. Given that

1. $f_{\lambda_0}(0) = 0$,
2. $f_{\lambda_0}'(0) = 1$,
3. $f_{\lambda_0}''(0) \neq 0$, and
4. $\partial f_\lambda / \partial \lambda \,|_{\lambda=\lambda_0} \neq 0$.

Then there exists an interval I about 0, and a smooth function $h: I \to R$ satisfying $h(0) = \lambda_0$, such that $f_{h(x)}(x) = x$. Moreover, $h'(0) = 0$ and $h''(0) \neq 0$.

The tangent bifurcation has received considerable attention of engineers recently because it is known that a beam or small arch bending under a load undergoes a tangent bifurcation at its buckling point. Additionally, astronomers have identified the equilibrium response of massive cold stars with the tangent bifurcation.

A bifurcation familiar to us from the analysis of the logistic equation is the **period-doubling bifurcation.** We now give another example of the period-doubling bifurcation. Consider the family of functions $f_\lambda(x) = \lambda e^x$, for $\lambda < 0$. Graphs of these functions are shown in Figure 4.26. When $\lambda = -e$, $f_\lambda(-1) = -1$, and $|f_\lambda'(-1)| = |-e(e^{-1})| = |-1| = 1$; thus $p = -1$ is a non-hyperbolic fixed point. We would expect a bifurcation to occur as λ passes through the critical value $-e$, and indeed one does occur. When $\lambda > -e$, the fixed point p is attracting, and when $\lambda < -e$, it is repelling. Hence, the nature of the fixed point changes as λ passes through $-e$, but this is not all that occurs. When $\lambda < -e$, the graph of $f_\lambda^2(x)$ is an increasing function that is concave up if $f_\lambda(x) < -1$, and concave down if $f_\lambda(x) > 1$. Since $f_\lambda^2(x)$ has two fixed points, this corresponds to a two-cycle in $f_\lambda(x)$. Figure 4.27 shows the bifurcation diagram for the above example, again plotting the fixed point and the periodic points on the y-axis against the parameter λ.

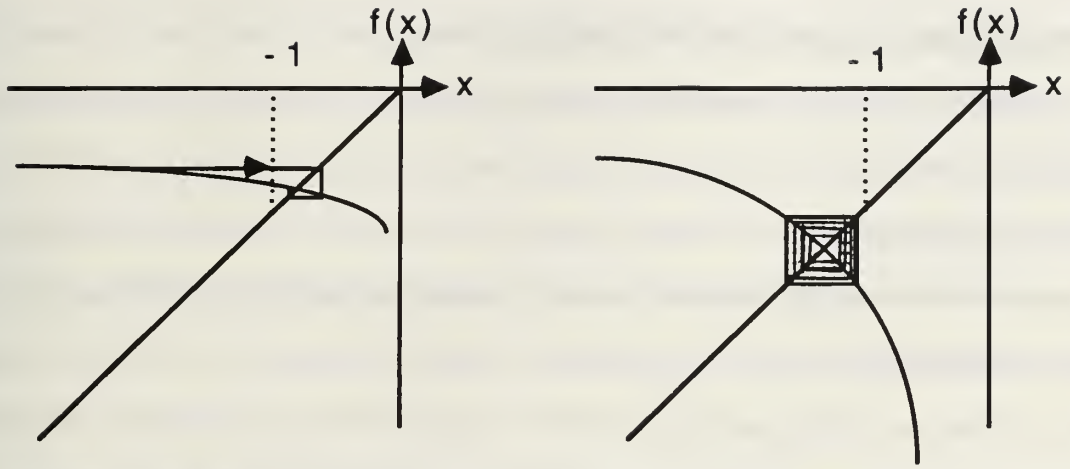

Figure 4.26 Graphs of the family $f_\lambda(x) = \lambda e^x$ for $\lambda > -e$ and $\lambda < -e$.

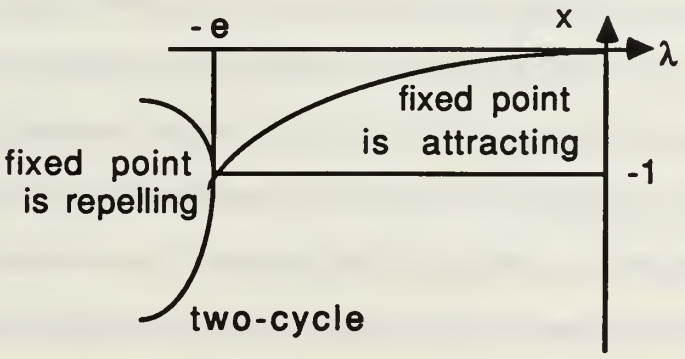

Figure 4.27 Bifurcation diagram for $f(x) = \lambda e^x$ with $\lambda < 0$.

In a period-doubling bifurcation, as the parameter passes through its critical value, the attracting fixed point becomes repelling and a cycle of period two emerges. Of course, as we saw with the logistic equation, a bifurcation can also occur that makes an attracting n-cycle repelling and creates an attracting 2n-cycle; hence the name "period-doubling." The formal mathematical characterization of the period-doubling bifurcation is given in the following result.

THE PERIOD-DOUBLING BIFURCATION. Given that

1. $f_\lambda(0) = 0$ for all λ in an interval about λ_0,
2. $f_{\lambda_0}'(0) = -1$, and
3. $\partial(f_\lambda^2)'/\partial\lambda \mid_{\lambda=\lambda_0}(0) \neq 0$.

Then there is an interval I about 0, and a function $h: \mathbf{I} \rightarrow \mathbf{R}$, such that $f_{h(x)}(x) \neq x$, but $f_{h(x)}^2(x) = x$.

As shown earlier, without loss of generality we are able to consider functions with fixed points at the origin and apply the same results in the general case for arbitrary fixed points. The proofs of both results concerning bifurcations can be found in Devaney (1989), and involve little more than application of the Implicit Function Theorem and knowledge of partial derivatives.

A situation where period-doubling bifurcations have been observed occurs in the field of electrocardiology. Electrochemical events in the heart, monitored by electrocardiograms (ECG), show periodic activity within the atria and ventricles of the heart. Abnormal cardiac rhythms, such as arrhythmia, have long been referred to as "chaotic heart action" (in the descriptive, non-mathematical sense) by cardiologists. Mostly, these *fibrillations*, or chaotic heart rhythms, have been recorded on ECGs immediately prior to a patient's death. By concentrating on the heart's natural pacemakers, called *ectopic foci* and which are located throughout the heart, cardiologists have found that in arrhythmia, the rhythm of the heart undergoes a series of period-doubling bifurcations leading eventually to chaotic behavior (in the precise mathematical sense). While the human heart requires a much more complex model than the one-dimensional

dynamical system, there is a high correlation between the period-doubling we have studied and the period-doubling on the ECG. Cardiologists are using this new theory to try to explain what actually happens in the heart in order to prevent or cure the condition. A current solution to this problem is to identify patients who are susceptible to arrhythmia and implant in their chests a device that detects fibrillations and gives the heart an electronic "kick" out of its period-doubling path. This device has proven successful, but it has been difficult to identify the patients who will benefit from its use.

Turbulence is one of the classical and unexplained phenomena in physics. In the past it has been so inaccessible to physicists and engineers that systems have been designed normally with "fudge" factors, or factors of safety, designed to compensate for the effects of turbulence. However, the issue of turbulence has rarely been directly taken on. Recently, in a physical model that creates turbulence in a very simple fluid setting, scientists have observed period-doubling bifurcations in the fluid leading directly to the onset of turbulence. These observations may be the first steps towards an understanding of this elusive behavior.

A few more examples of how a changing parameter in the real world can lead to bifurcations will emphasize the importance of studying this concept. In the logistic equation it is reasonable to assume that the parameter λ, which measures the species' interaction with its environment, does not change too drastically from cycle to cycle. On the other hand, in a fluid system, it is just as reasonable to see how the changing velocity of a flow can change a parameter quickly enough to induce the period-doubling bifurcations that lead to turbulence. So it might seem reasonable to surmise that the panel

flutter on a supersonic aircraft could be avoided through a careful design that stays away from dangerous parameter values. However, not every atmospheric condition can be predicted accurately by a model, or duplicated in a wind tunnel. Moreover, atmospheric phenomena, such as ice buildup on wings or wind shears, can quickly push parameters into critical regions. In complex systems one can rarely anticipate every set of parameter values, which is an important reason to pay careful attention to parameter space.

J. SARKOVSKII'S THEOREM

We now present another remarkable result about one-dimensional dynamical systems due to the Russian mathematician, A. N. Sarkovskii. The development here is from the article by Devaney (1989). First, order the positive integers in the following manner:

$$3 \gg 5 \gg 7 \gg 9 \gg \cdots \gg$$

$$2\cdot 3 \gg 2\cdot 5 \gg 2\cdot 7 \gg 2\cdot 9 \gg \cdots \gg$$

$$2^2\cdot 3 \gg 2^2\cdot 5 \gg 2^2\cdot 7 \gg 2^2\cdot 9 \gg \cdots \gg$$

$$2^3\cdot 3 \gg 2^3\cdot 5 \gg 2^3\cdot 7 \gg 2^3\cdot 9 \gg \cdots \gg$$

$$\cdots \gg$$

$$\gg \cdots \gg 2^4 \gg 2^3 \gg 2^2 \gg 2 \gg 1.$$

Here, we have first listed the odd numbers in ascending order, followed by two times the odd numbers, then 2^2 times the odd numbers and so on, through every positive integer power of 2 times the odd numbers, finally followed by the powers of 2 in descending order. Given this ordering of the positive integers, we present Sarkovskii's theorem.

SARKOVSKII'S THEOREM. Suppose $f: \mathbf{R} \to \mathbf{R}$ is continuous. If f has a periodic point of period n, then f also has a periodic point of period k for all k with $n \gg k$ in the above ordering.

While this result only holds for one-dimensional systems, it is remarkably powerful due to its lack of hypothesis. In fact, the only requirement on f is that it be continuous.

Rather than proving Sarkovskii's theorem itself, we prove a special case in what follows. This proof is found in the article by Devaney (1989) and differs from the complete proof of Sarkovskii's theorem by requiring less bookkeeping. The full proof of Sarkovskii's theorem can be found in Devaney (1989).

PROPOSITION. Suppose a continuous map $f: \mathbf{R} \to \mathbf{R}$ has a cycle of period 3. Then f has periodic points of all periods.

PROOF. The proof depends on two observations. First, if I and J are closed intervals with $J \supset I$ and $f(I) \supset J$, then f has a fixed point on I. (This is similar to the result about fixed points we proved in Section IIC). The second observation is as follows: suppose $A_0, A_1, \ldots, A_n$ are closed intervals such that $f(A_i) \supset A_{i+1}$ for $i = 0, 1, \ldots, n\text{-}1$. Then there exists at least one subinterval J_0 of A_0 which is mapped onto A_1. Similarly, there is a subinterval J_1 of A_1 mapped onto A_2, and hence a subinterval J_1 of J_0 such that $A_1 \supset f(J_1)$ and $f^2(J_1) = A_2$. Continuing in this manner, we obtain a nested sequence of subintervals which map into the various A_i, in order.

To prove the proposition, let a, b, and c be such that $f(a) = b$, $f(b) = c$, and $f(c) = a$. With only a slight loss of generality, assume $a > b > c$ (the other case is handled similarly). This hypothesis is shown in Figure 4.28.

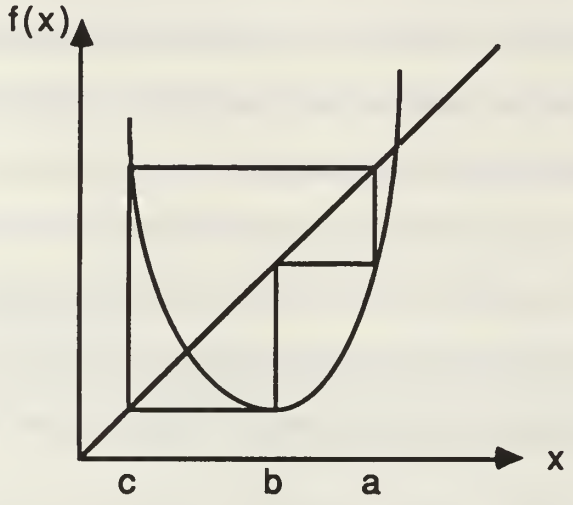

Figure 4.28 A three-cycle with $f(a) = b$, $f(b) = c$, $f(c) = a$, and $a > b > c$.

We let $I_0 = [b, a]$ and $I_1 = [c, b]$. By our assumptions, $f(I_0) \supset I_1$ and $f(I_1) \supset I_0 \cup I_1$. Figure 4.27 shows there is a fixed point between c and b. Similarly, f^2 has fixed points between a and b, so at least one of these points must be of period 2. Fixing $n > 3$, we now produce a cycle of period n.

We first find a nested sequence $A_0, A_1, \ldots, A_{n-2}$, of subintervals of I_1 as follows: let $A_0 = I_1$. Since $f(I_1) \supset I_1$, there is a subinterval A_1 of A_0 such that $f(A_1) = A_0 = I_1$. By induction, we can find a subinterval A_{n-2} of A_{n-3} such that $f(A_{n-2}) = A_{n-3}$, $f^2(A_{n-2}) = A_{n-4}, \ldots,$ and $f^{n-2}(A_{n-2}) = A_0 = I_1$. Since $f(I_1) \supset I_0$, there is a subinterval A_{n-1} of A_{n-2} such that $f^{n-1}(A_{n-1}) = I_0$. Finally, since $f(I_0) \supset I_1$, we have $f^n(A_{n-1}) \supset I_1 \supset A_{n-1}$. Hence, f^n has a fixed point in A_{n-1} from the first observation. We now show this point has period n under

f. Since $I_1 \supset f^i(A_{n-1})$ for $i = 0, 1, \ldots, n-2$, but $I_0 \supset f^{n-1}(A_{n-1})$, and $f^n(A_n) \supset I_1$, this point has its first n-2 iterates in I_1, then jumps to I_0, in the n - 1 iteration, and finally back to I_1. This completes the proof.

In addition to the above proposition, another corollary to Sarkovskii's theorem states: *If f has a periodic point which is not a power of 2, then f has infinitely many periodic points.*

Sarkovskii's theorem provides considerable information about a function. For example, it would be very difficult to check directly whether the function $f(x) = 1 + (5/2)x - (3/2)x^2$ is chaotic. However, since $f(0) = 1$, $f(1) = 2$, and $f(2) = 0$, the function has a three-cycle. Thus Sarkovskii's theorem tells us it has cycles of all periods. Hence, we automatically know it is chaotic without having to check the three defining conditions.

If we consider the period-doubling bifurcations as a route to chaos, then only finitely many periodic points must have the periods $1, 2, 2^2, 2^3, 2^4, \ldots, 2^N$ for some N. Then as the parameter varies and the dynamics of the system become more complex, we introduce periods in a specific order 2^{N+1}, $2^{N+2}, \ldots$. This argument does not claim that the new orbits appear as period-doubling bifurcations, but that something similar must occur.

We cannot derive a converse to the theorem from the special ordering. If we find a cycle of period k, and $n \gg k$ in the ordering, there is no guarantee that a cycle of period n exists.

While we used a graphical analysis to show that the Baker map $f(x) = 2x \pmod 1$ is chaotic, we now confirm this using Sarkovskii's theorem. Although the Baker map is not continuous, since every iterate has fixed

points on [0, 1], we can apply the above proof of the proposition directly to it. Thus, if we can find a three-cycle of the Baker map, we will know it is chaotic. But that is easy: since the points 1/7, 2/7, and 4/7 form such a cycle, the Baker map is chaotic.

Another example using Sarkovskii's theorem to find chaotic behavior is with the function $f(x) = x^2 + c$, where $c = -1.755$. Using a computer, we can verify the attracting orbit of period 3 given by $f(0) = -1.755$, $f(-1.755) \approx 1.325$ (to four decimal places), and $f(1.325) = 0$. This three-cycle guarantees cycles of all periods and an infinite number of periodic points. Nevertheless, regardless of the initial value, the orbit is always attracted to this three-cycle. But then where are the other periodic points if only three are found by computational iteration? The answer is that all other cycles are repelling and, because of computer round-off error, iterates always move away from a repelling cycle (unless the points are rational numbers represented exactly up to the precision of the computer). This example further illustrates the sensitive dependence on initial conditions of chaotic dynamical systems.

Now why, of all the infinite orbits, is only *one* of them attracting? This question is extremely complicated in general. However, in the case of our specific example it can be answered using a result from complex analysis. For a complex analytic map, each attracting orbit attracts at least one critical point for the map. Now the map $f(z) = z^2 - 1.755$, is an analytic function with only one critical point $z = 0$. Hence the function has only one attracting orbit.

K. THE QUADRATIC FAMILY REVISITED

The last example showed another member of the quadratic family of maps; namely, $f(x) = x^2 + c$. We now study this map in greater detail as we vary the parameter c. The results in this section are from Falconer (1990) and the article by Devaney (1989).

First note that for $c > 1/4$ the graph of $f(x) = x^2 + c$ lies above the diagonal $y = x$ and $f^n(x) \to \infty$ for all $x \in \mathbf{R}$. For $c = 1/4$, the graph of $f(x)$ is tangent to the diagonal when $x = 1/2$ (which is a single fixed point). Finally, for $c < 1/4$, f has two fixed points which we denote by p_1 and p_2 where $p_1 < p_2$. This is an example of the tangent bifurcation as c passes through the value 1/4. These three cases are shown in Figure 4.29.

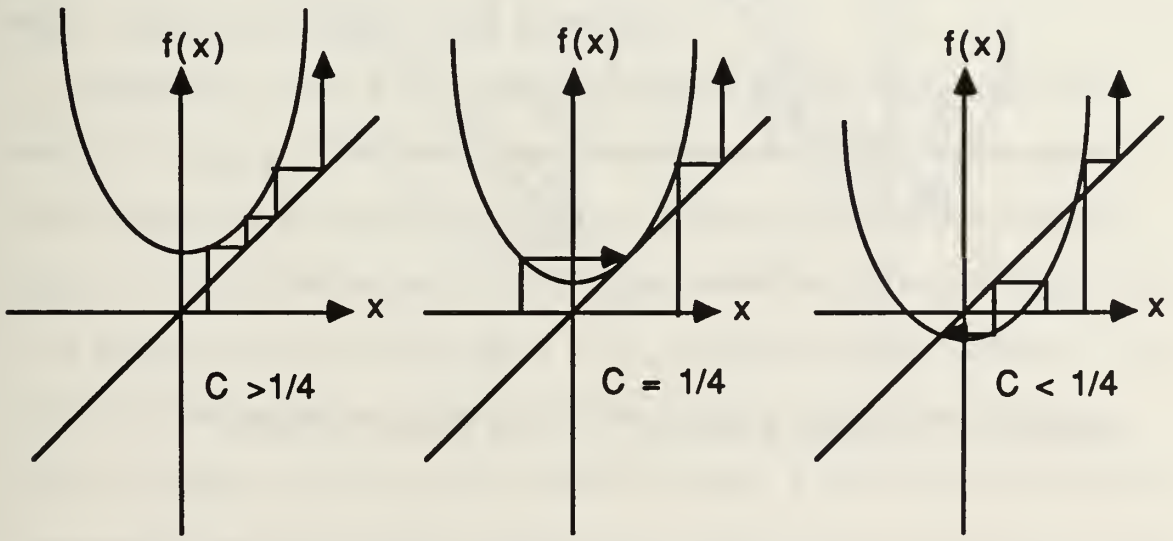

Figure 4.29 Graphs of $f(x) = x^2 + c$ for $c > 1/4$, $c = 1/4$, and $c < 1/4$.

Now observe that for all $c < 1/4$, $p_2 > 1/2$, so $|f'(p_2)| = |2p_2| > 1$. However, for $-3/4 < c < 1/4$, $|f'(p_1)| < 1$ since $-1/2 < p_1 < 1/2$. Finally for $c < -3/4$, $|f'(p_1)| > 1$ since $p_1 < -1/2$. This demonstrates that after the tangent bifurcation occurs (as c passes through 1/4) the fixed point p_1 is attracting and p_2 is repelling. However, as c passes through -3/4, p_1 also becomes repelling (and in fact, a two-cycle forms around p_1). The graphs of $f(x) = x^2 + c$ as c passes through -3/4 are shown in Figure 4.30. As c continues to decrease, we get another sequence of period-doubling bifurcations, similar to those we saw with the logistic equation. The frequency of these bifurcations is also governed by the Feigenbaum constant.

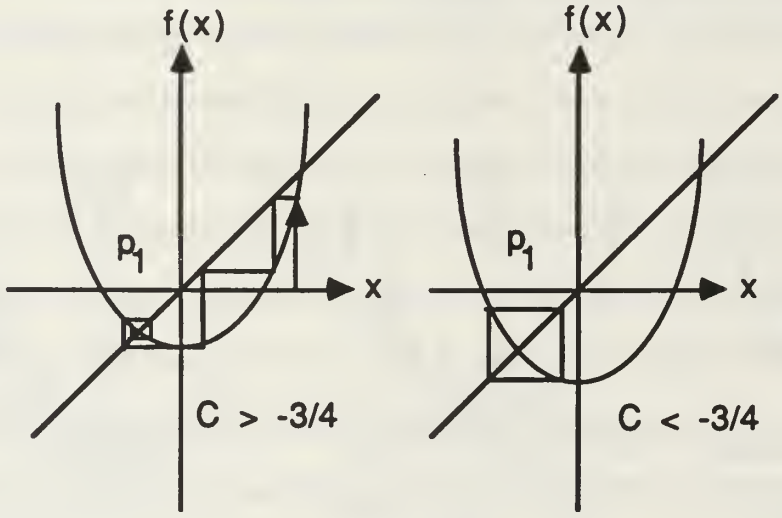

Figure 4.30 The graph of $f(x) = x^2 + c$ as c passes through -3/4.

We have developed three ways to determine if this period-doubling leads to chaos. The first is to graph f(x), and $f^2(x)$ through $f^n(x)$, for certain values of c and observe the fixed points of these graphs. The second way is to solve algebraically for the roots of $f^n(x) = x$ and observe the fixed points as the roots

of these polynomials. For the case f^2, solving the equation $f^2(x) = x$ yields the fourth-order equation $(x^2 + c)^2 + c = x$, or $x^4 + 2cx^2 - x + c^2 + c = 0$. Since we already know two of the roots (p_1 and p_2) for any given value of c, we can solve directly for the other two. However, as n increases, it becomes very difficult to solve for the roots of $f^n(x)$. Third, we can simply find a three-cycle for certain values of c (for example $c = -1.755$) and appeal to Sarkovskii's theorem as proof that periods of all other orders do exist.

Continuing to analyze the family $f(x) = x^2 + c$, we see from Figure 4.28 that for all $c < 1/4$, if $|x_0| > p_2$, then $f^n(x_0) \to \infty$. Thus, we can focus our attention on the interval $I = [-p_2, p_2]$ where all of the interesting dynamics occur. Let us further restrict our attention to the range of parameter values $c < -2$. If we analyze this function on a computer for almost any initial value, the iterates of f to go to infinity. However, as shown below, there are many orbits which do not escape under iteration of f.

The graph of f for $c < -2$ is shown in Figure 4.31. If we consider the interval $I = [-p_2, p_2]$ note that there is a subinterval A_0 of I that maps to values outside of I, hence all points in A_0 will escape to ∞. Consider the graph of f^2, a similar analysis to the one performed for the logistic map with $\lambda > 4$ shows that two subintervals of $I - A_0$ are again mapped outside of I, so again points escape to ∞. The second iteration of f on the interval I is shown in Figure 4.32. The subintervals that escape I on the second iteration are labeled A_1 and A_2 in the figure.

Analyzing the points that remain in I after infinite iteration, we deduce a Cantor set has been constructed on I. While it is not necessarily the classical Cantor set, nevertheless, it contains no intervals despite an infinite number

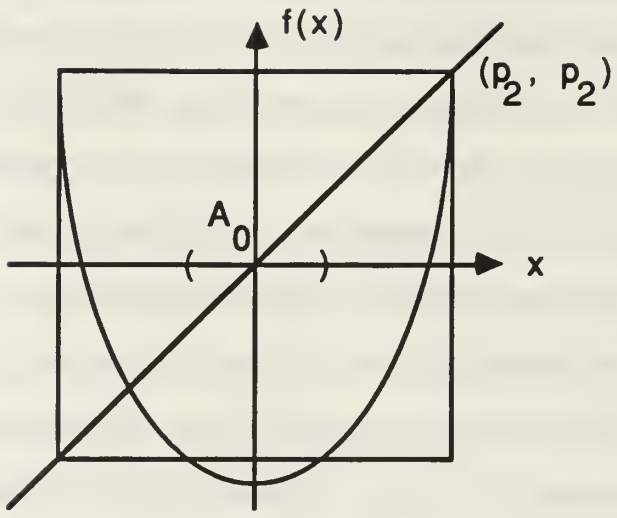

Figure 4.31 The graph of $f(x) = x^2 + c$ for $c < -2$.

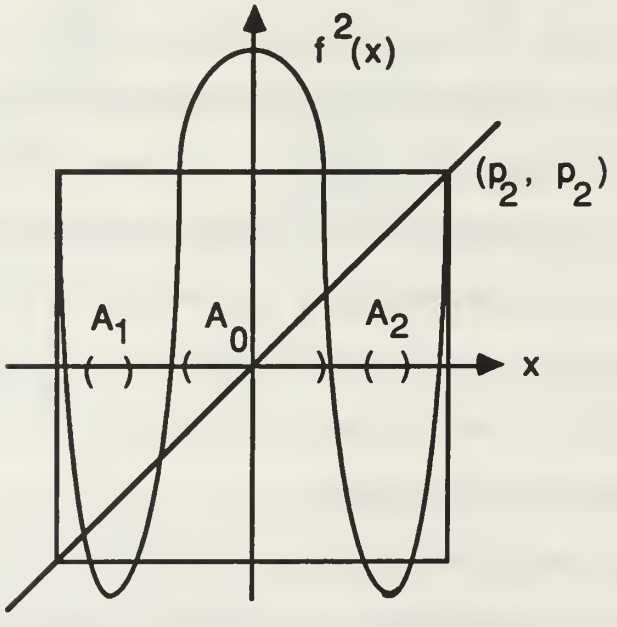

Figure 4.32 The graph of $f^2(x)$ for $c < -2$.

of points. (As we mentioned in Section IIE, this could be an example of a "fat" Cantor set, depending on the value of c). We now define the set

$$\Lambda = \{x \in I: f^n(x) \in I \ \forall \ n \geq 0\},$$

and assert that Λ is a Cantor set.

The dynamics of f on $R - \Lambda$ are quite simple because every initial value tends to $+\infty$ under infinite iteration. We want to know what happens to the set Λ. To determine this, we simplify the analysis through symbolic dynamics. Recall that the shift map on code space $\sigma: \Sigma \to \Sigma$ is a continuous mapping. We now try to relate σ and f. When we remove the interval A_0 from I, two subintervals remain, denoted by I_0 and I_1 (see Figure 4.33). Hence, if $x \in \Lambda$, then $f^n(x) \in I_0 \cup I_1$ for all $n \geq 0$. Next define the **itinerary** of x by $S(x) = (s_0 s_1 s_2 \ldots)$ where $s_i \in \{0, 1\}$ and $s_i = k$ if and only if $f_i(x) \in I_k$.

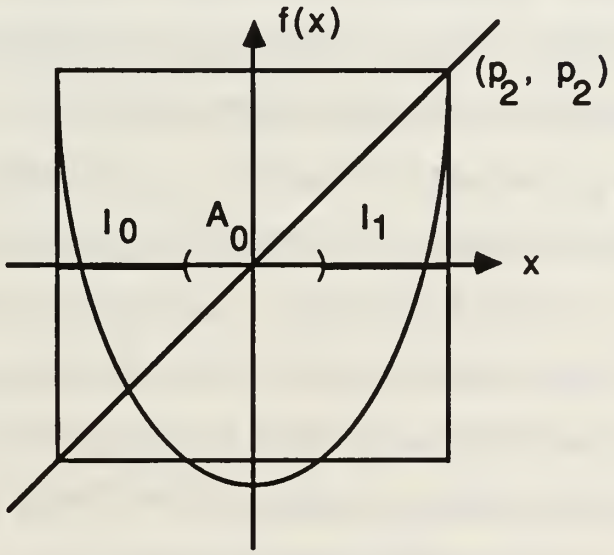

Figure 4.33 The subintervals I_0 and I_1.

We now show that $S: \Lambda \to \Sigma$ is a homeomorphism. First, to see that S is one-to-one, let $x, y \in \Lambda$ and suppose $S(x) = S(y)$. Then for each n, $f^n(x)$ and $f^n(y)$ both lie on the same side of A_0. It follows that f is monotonic on the interval between $f^n(x)$ and $f^n(y)$. Hence, all points in this interval remain in $I_0 \cup I_1$. This observation contradicts the fact that Λ is a Cantor set and contains no intervals, so S is one-to one.

To see that S is onto, for a closed interval J, set $f^{-n}(J) = \{x \in I: f^n(x) \in J\}$. If $I \supset J$, then $f^{-1}(J)$ consists of two subintervals, one in I_0 and one in I_1. Let $s \in \Sigma$ with $s = s_0 s_1 s_2 \ldots$, and define

$$I_{s0\ldots sn} = \{x \in I: x \in I_{s0}, f(x) \in I_{s1}, \ldots, f^n(x) \in I_{sn}\},$$

so $I_{s0\ldots sn} = I_{s0} \cap f^{-1}(I_{s1}) \cap \ldots \cap f^{-n}(I_{sn})$. We claim that the $I_{s0\ldots sn}$ form a nested sequence of nonempty closed intervals as $n \to \infty$. Note that $I_{s0\ldots sn} = I_{s0} \cap f^{-1}(I_{s1\ldots sn})$. By induction, we assume that $I_{s1\ldots sn}$ is a nonempty closed subinterval so that $f^{-1}(I_{s1\ldots sn})$ consists of two subintervals, one in I_0 and one in I_1. Hence, $I_{s0\ldots sn}$ is a single closed subinterval. These intervals are nested since $I_{s0\ldots sn} = I_{s0\ldots sn-1} \cap f^{-1}(I_{sn}) \subset I_{s0\ldots sn-1}$. Hence, the intersection $\cap_{n\geq 0} I_{s0\ldots sn}$ is nonempty for any $n \geq 0$. Note that if $x \in \cap_{n\geq 0} I_{s0\ldots sn}$, then $x \in I_{s0}$, $f(x) \in I_{s1}, \ldots$, so that $S(x) = s_0 s_1 s_2 \ldots$. This proves that S is onto.

Since S is one-to-one and onto, it follows that S^{-1} exists. Hence we need only show that S is continuous to prove it is a homeomorphism. Thus, choose $x \in \Lambda$ and let $S(x) = s_0 s_1 s_2 \ldots$. Let $\varepsilon > 0$, and choose n such that $1/2^n < \varepsilon$. Consider $I_{t0\ldots tn}$ for all possible combinations of $t_0, t_1, \ldots, t_n$. These subintervals are disjoint, and Λ is contained in their union. Choose $y \in \Lambda$ and δ such that $|x - y| < \delta$. Then $y \in I_{s0\ldots sn}$, and $S(x)$ and $S(y)$ agree for the first $n+1$ terms. Hence, by the metric on code space Σ_2, $d(S(x), S(y)) < 1/2^n < \varepsilon$.

Thus S is continuous. Trivially, S^{-1} is also continuous; hence, S is a homeomorphism.

Since $S: \Lambda \to \Sigma$ is a homeomorphism , we use topological conjugacy to show that f has the same dynamics as the shift map σ on code space Σ_2. The commutative diagram for this relationship is shown in Figure 4.34. Since we know σ is chaotic on Σ, f is chaotic on Λ through topological conjugacy.

$$\begin{array}{ccc} \Lambda & \xrightarrow{\ f\ } & \Lambda \\ \downarrow S & & \downarrow S \\ \Sigma & \xrightarrow{\ \sigma\ } & \Sigma \end{array}$$

Figure 4.34 The topological conjugacy between $f: \Lambda \to \Lambda$ and $\sigma: \Sigma \to \Sigma$.

We now return to the map $f(x) = x^2 - 1.755$ to learn more about it through its symbolic dynamics. Since it has a three-cycle, -1.755, 0, and 1.325. . . , Sarkovskii's theorem guarantees it is chaotic. However, we see we could have discovered this feature without Sarkovskii's theorem.

We start by finding three open intervals, O_1 about 0, O_2 about -1.755, and O_3 about 1.325. Select these such that O_i contains the closure of $f(O_i)$, and $O_{i+1} \supset f(O_i)$. We can always make this choice because f is a continuous map (although in practice, the use of a computer would help). Now let I_0 denote the closed interval between O_1 and O_3 and let I_1 be the closed interval between O_2 and O_1. This relationship is shown schematically in Figure 4.35. We may choose each O_i such that $|(f^3)'(x)| > 1$ on $I_0 \cup I_1$. We then have $f(I_0) \supset I_1$ and $f(I_1) \supset I_0 \cup I_1$, so each interval is stretched over its image.

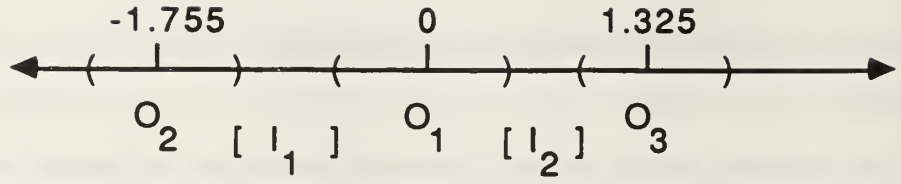

Figure 4.35 The closed intervals I and the open intervals O.

We now introduce symbolic dynamics. Let $\Lambda = \{x: f^n(x) \in I_0 \cup I_1 \; \forall \; n \geq 0\}$. We know that Λ is a Cantor set. To model the dynamics of f on Λ, consider modified code space Σ' where

$$\Sigma' = \{s_1 s_2 s_3 \ldots : s_i \in \{0, 1\} \text{ and } s_k = 0 \Rightarrow s_{k+1} = 1\},$$

i.e., this is just Σ_2 with no adjacent pairs of 0s. If we now define the map $S: \Lambda \rightarrow \Sigma'$ as above, we see that the condition $f(I_0) \supset I_1$ forces the condition of no adjacent 0s in Σ'. The diagram in Figure 4.36 shows how f commutes with σ through S. Hence the shift map $\sigma: \Sigma' \rightarrow \Sigma'$ provides all of the information about the dynamics of $f: \Lambda \rightarrow \Lambda$. Thus there exist points of all periods in Σ'. In fact, the point 0111...10111...10..., with blocks of n-1 repeating 1s, is the same point found in the proof of the special case of Sarkovskii's theorem.

$$\begin{array}{ccc} \Lambda & \text{— f} \rightarrow & \Lambda \\ \downarrow S & & \downarrow S \\ \Sigma' & \text{— } \sigma \rightarrow & \Sigma' \end{array}$$

Figure 4.36 Topological conjugacy between f and σ.

L. JULIA SETS

Julia sets, along with the Mandelbrot set, have perhaps been the most significant factors in generating interest in chaos among laymen. The reason is because the intricate and beautifully colored computer images shown as "pictures of chaos" are normally pictures of Julia sets. Moreover, the "movies" of these images, exploding across the screen, are simply the Julia sets viewed under the continuous changing of a parameter.

Julia sets were actually discovered in the 1920s by the French mathematicians Gaston Julia and Pierre Fatou. However, their true beauty and intricate detail were not fully realized until the 1970s when computer graphics allowed for their inspection in detail. The concept of a Julia set can be understood with only a basic understanding of complex numbers. On the other hand, a formal and mathematically rigorous treatment of Julia sets requires a theory of complex analysis beyond the scope of this thesis. Here we present only a cursory survey of Julia sets in their ambient space, the complex plane, still treating one-dimensional maps in the iterated function systems. While many of the references cited discuss Julia sets, the presentation here is from Keen (1989) and Falconer (1990). There are many equivalent definitions of Julia sets, but the one we present is perhaps the simplest to demonstrate and understand.

DEFINITION. Given a mapping $f: \mathbf{C} \to \mathbf{C}$ of the complex plane, its Julia set $J(f)$ is the closure of the set of repelling periodic points of f.

As a simple example, consider the map $f(z) = 2z$. Under infinite iteration of f, all points in the complex plane excluding the origin tend to ∞ (or to be more precise, the point at ∞). The origin is a fixed, and hence periodic point of this map. Since iteration of any point other than the origin tends away from it, the origin is repelling. This result is also verified from $|f'(0)| = |2| > 1$. Since the only repelling periodic point of this map is the origin, which is its own closure, the Julia set for $f(z) = 2z$ is the origin.

We now present a less trivial example which demonstrates many interesting properties of Julia sets. Consider the map $f(z) = z^2 + c$ for $c = 0$. All points inside the unit circle $|z| < 1$ tend to the origin under infinite iteration. Thus the origin is an attracting fixed point of the map. In fact, $|f'(0)| = |2(0)| = 0$, which also verifies that the origin is attracting. Moreover, all points $|z| > 1$ outside the unit circle tend to ∞ under iteration of this map.

Now consider the standard unit circle, $|z| = 1$. These points are represented by $z = e^{i\theta}$. Then $z^2 = e^{i2\theta}$, which is exactly the chaotic map of the unit circle $f: S^1 \rightarrow S^1$, where $f(\theta) = 2\theta$, studied earlier. We know the periodic points of this map are dense on the unit circle. Since the periodic points of $f(z) = z^2$ are dense on the unit circle, every point on the unit circle is the limit of a sequence of periodic points of f. Thus, the closure of the periodic points of f is the unit circle. Moreover these points are repelling. To see this, recall that points inside the unit circle converge to the origin, whereas points outside the unit circle diverge to the point at infinity. Furthermore, $|f(z)|\,|_{|z|=1} = |2(1)| = 2 > 1$, verifying that these points are repelling. Hence, the Julia set for the function $f(z) = z^2$ is the standard unit circle.

Another concept associated with a Julia set is that of the filled Julia set, denoted F(f). When the Julia set is a closed curve, the set F(f) is the union of J(f) with its interior. The filled Julia set is the set of points that do not escape to ∞ under infinite iteration of f. For the example $f(z) = z^2$, the filled Julia set is the closed unit disc, $F(f) = \{z: |z| \leq 1\}$. As an aside, the complement of the Julia set is called the **Fatou set** and is sometimes denoted F(f) as well, although J^c is also used. Loosely speaking, J(f) is the set containing the "bad" (i.e., chaotic) behavior, while the Fatou set is the "good" set, possesing the well-behaved dynamics.

Having introduced the concepts of Julia sets and filled Julia sets in this simple setting, we now describe an algorithm for generating computer images of these objects. If we superimpose the complex axes on a computer screen to an appropriate scale, then points in the complex plane correspond to pixels on the screen, although this relationship is certainly not one-to-one. Given a function f(z), we can iterate each pixel. Since we are interested in the points that escape to ∞, a bound (normally very large) can be set which we call |Z|, and above which an iterate is considered as having escaped. Next select k integers $N_1 < N_2 < \ldots < N_k = N$. Color the screen with k + 1 colors based on the following algorithm: as a point is iterated, if it has not escaped after N iterations, color it black. If it escapes (goes beyond |Z|) between 0 and N_1 iterations, assign to it another color (say, red). If it escapes between N_1 and N_2 iterations, color it with yet another color (for example, yellow). Continue in this manner until the entire screen has been colored. Selecting a large value for k provides more detail, which can be refined further by experimentally adjusting N and |Z| with respect to each other. The part of

the screen colored black (if we have chosen N, |Z|, and the scale appropriately) is the filled Julia set for f, and this region's boundary (provided it is connected) is the Julia set itself. The colored bands around the Julia set are contours corresponding to the various escape times of points in the exterior of the Julia set. The reason the complement of the filled Julia set is colored is because of the finite scale of the computer screen: there can be great detail occurring within the area of a single pixel and, while the complete Julia set is not revealed by just the black area, much can be determined about its border by examining the distorted contours surrounding it.

For the example $f(z) = z^2$, coloring the screen based on this escape time algorithm produces a black disc with a sequence of concentric colored circles around it (which in itself is not particularly interesting). However, recalling the family of functions $f_c(z) = z^2 + c$, as the parameter c is varied some very interesting results occur. Unfortunately, while this Julia set has many fascinating properties, an advanced level of complex analysis is required to establish even its most basic properties. The required concepts include families of normal functions, the Arzela-Ascoli theorem, and Montel's theorem. These results are beyond the scope of this thesis, but an excellent summary of them is found in Falconer (1990), and we provide a synopsis of them at the end of this section. Nevertheless, we can still provide a brief description of some of the salient characteristics.

As the parameter c is varied away from the origin, the Julia set (the unit circle) begins to continuously distort and take on different shapes. Closer inspection reveals that the boundary appears to become infinitely detailed and self-similar; in fact, it becomes fractal. (Even with c = 0, the boundary of

the filled Julia set is infinitely self-similar, although in a trivial manner.) While we saw that the dynamics of f on $J(f)$ for $c = 0$ are chaotic (since they share the properties of $f(\theta) = 2\theta$) they continue to be chaotic on $J(f)$ as c varies. For the quadratic family $f(z) = z^2 + c$, as c is varied the Julia set varies from a circle, to a closed curve with "bulbs" that are "pinched" together at a single point, to "dendrites" which are fractal structures with no interior, to "dust" which is a set of disconnected points which are scattered about a region of the complex plane, similar to the Cantor set. For some Julia sets with fractal boundaries, like the Koch snowflake, the lengths of the boundaries are infinite. A further result about Julia sets is that they are either **connected** (meaning they consist of one solid piece) or **totally disconnected** (meaning they have a structure similar to Cantor dust). The Julia sets for various values of c are shown in Figure 4.37.

Without going into too much detail, we provide a brief synopsis of the most important ingredients of the mathematical theory behind these results. For analytic functions in $\mathbf{C}$ (i.e., those that are infinitely differentiable in the complex sense) techniques of complex variable theory can be used to establish the basic properties of Julia sets.

It can be shown that an alternative (but equivalent) way of defining the Julia set $J(f)$ for polynomials f is as the set of all complex values z for which the family $\{f^k(z)\}$ $k = 1, 2, \ldots$ is not normal. Loosely speaking, a family of complex analytic functions is said to be **normal** if it possesses some expecially strong convergence properties (called "uniform" convergence) on compact subsets of a given open set. Using a powerful result from complex analysis known as Montel's theorem, it is possible to demonstrate that if f is a

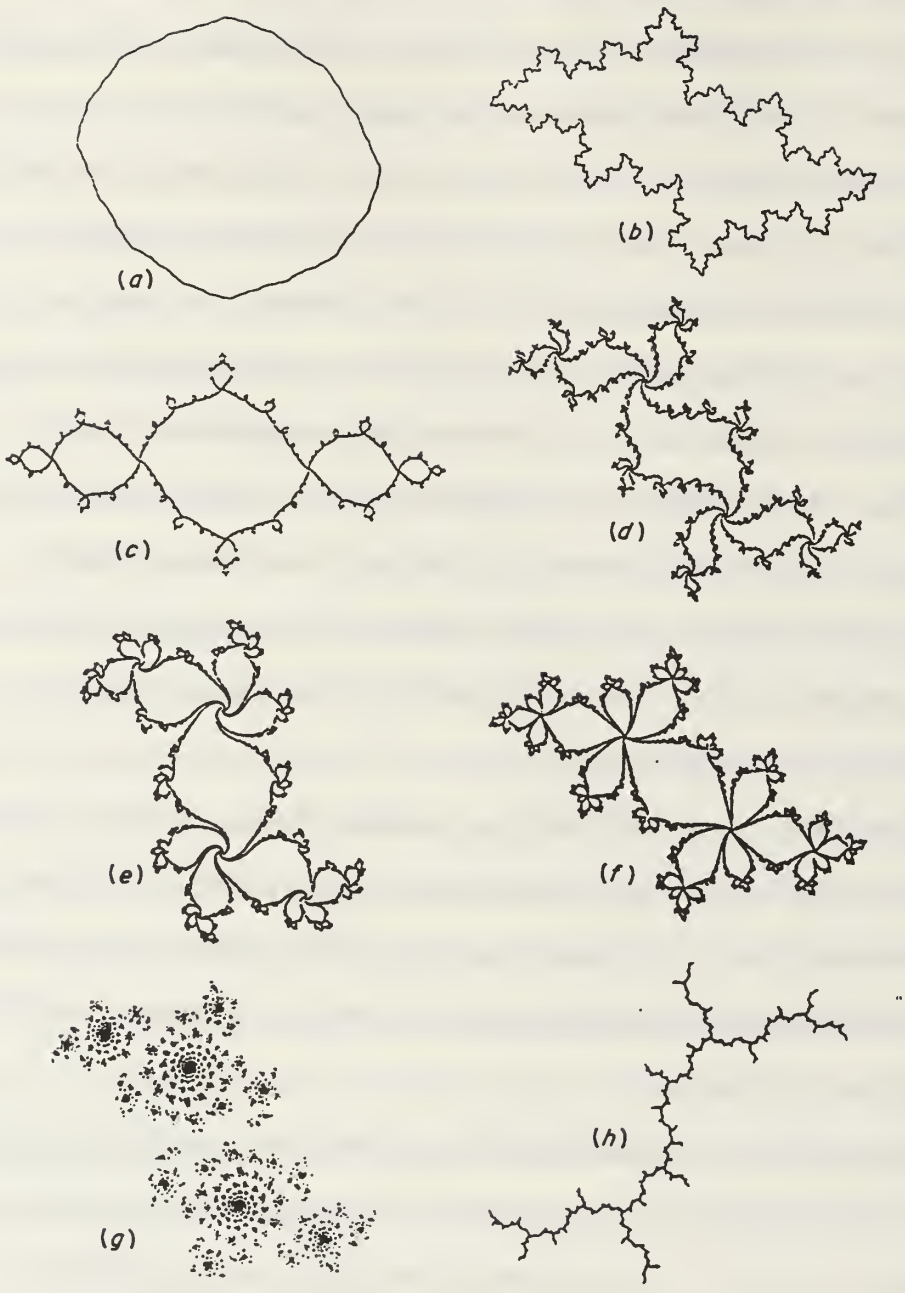

Figure 4.37 Julia sets of the function $f(z) = z^2 + c$ for (a) $c = -.1 + .1i$; (b) $c = -.5 + .5i$; (c) $c = -1 + .05i$; (d) $c = -.2 + .75i$; (e) $c = .25 + .52i$; (f) $c = -.5 + .55i$; (g) $c = .66i$; (h) $c = -i$. The figure is from Falconer (1990, p. 213).

polynomial, then J(f) is non-empty, compact (closed and bounded), contains no "isolated" points, and much more. Note that this is consistent with what we have seen with the quadratic family, even in the case when J(f) is "dust."

It should be pointed out that the above results do not necessarily hold for non-polynomial complex analytic maps. The Julia set for the exponential map $f(z) = e^z$, for example, is the entire complex plane. Of course, polynomial functions are not the only ones that generate interesting Julia sets. Some of the trigonometric families, such as $\lambda \sin z$, also provide very interesting characteristics as λ varies.

M. THE MANDELBROT SET

The Mandelbrot set is often associated with intricate computer graphics. It has been described from "the most complex object in mathematics" to "the most beautiful object in mathematics." While Julia sets are found in range space of a complex function, the Mandelbrot set lies in parameter space, which is the complex plane when the parameter is a complex number. Like the Julia set, almost every reference cited discusses the Mandelbrot set. However, th particular presentation here is based on Branner (1989) and Falconer (1990). There are two equivalent definitions of the Mandelbrot set, and both of them are presented here.

One definition of the Mandelbrot set for $f_c(z) = z^2 + c$ is the set of values of c for which the associated Julia set J(f) is connected. (This definition stresses the connection between the Mandelbrot set and Julia sets.) As mentioned in the last section, the Julia sets for $f(z) = z^2 + c$ vary from being totally connected to "dust." The values of c for which the Julia sets are dust

do not belong to the Mandelbrot set. An equivalent definition, which is perhaps easier to understand, follows.

DEFINITION. The **Mandelbrot set** $\mathcal{M}$ is the set of complex values of c for which the origin does not escape to ∞ under infinite iteration of $f(z) = z^2 + c$.

A picture of the Mandelbrot set in the complex plane is shown in Figure 4.38. The figure is from Falconer (1990, p. 205). We have already seen that for $c > 1/4$ on the real line, all values of x including the origin go to infinity under iteration of $f(x) = x^2 + c$; for $c < -2$, the origin also escapes. Hence, we know that the Mandelbrot set contains the interval $[-2, 1/4]$ on the real line. However, the situation is not nearly this simple when c varies in the complex plane.

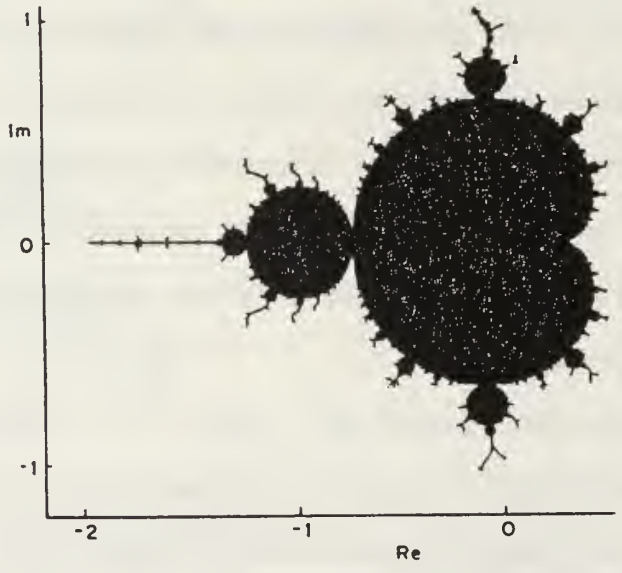

Figure 4.38 The Mandelbrot set.

Like Julia sets, the Mandelbrot set is infinitely detailed. In fact, it contains many smaller copies of itself around its border. However, it is by no means self-similar because it contains many other interesting shapes. Moreover, its intricate detail varies significantly among the border regions of the smaller, Mandelbrot-like sets.

Pictures of the Mandelbrot set can also be generated using the escape time algorithm used to draw Julia sets. Here the coloring of the complement becomes particularly important because many of the tendrils extending from the main body of the Mandelbrot set are too detailed to capture on a computer screen (regardless of the scale chosen), so they are only evidenced by the distorted contours surrounding them. It is known that the Mandelbrot set is connected: even points that appear isolated on computer images are connected to the main body by dendrites too small to be seen on a computer screen.

The first definition of the Mandelbrot set shows an intimate connection with Julia sets of the function $f(z) = z^2 + c$, but the Mandelbrot set contains even more information about the dynamics of the function f. Once again, however, any rigorous mathematical development of these dynamics requires the advanced theory of complex analysis. So again we only describe some of the more interesting characteristics.

The first result is that each "bulb" of the Mandelbrot set corresponds to an attracting k-cycle of $f(z)$ for a particular value of k. For example, the large central cardioid corresponds to the values of c for which $f(z) = z^2 + c$ has an attracting fixed point. To see this, note that an attracting fixed point must satisfy $z^2 + c = z$ and $|f'(z)| = |2z| < 1$. The boundary of this region is given

by $c = z - z^2$, where $|z| = 1/2$. In polar representation, this becomes $c = (1/2)\, e^{2\pi i\theta} - (1/4)\, e^{4\pi i\theta}$, $0 \le \theta < 2\pi$. These values of c trace out a cardioid in the complex plane with a cusp at $z = 1/4 + 0i$. Unfortunately, the periods of the attractive cycles of the other bulbs do not so easily reveal themselves mathematically.

It is perhaps not surprising that the periods of the bulbs along the real axis are in direct correspondence with the bifurcations found for the map $f(x) = x^2 + c$. Recall that for this latter map, a tangent bifurcation occurs at $c = 1/4$, and a series of period-doubling bifurcations begins as c decreases through $-3/4$. Figure 4.39 shows the Mandelbrot set plotted on the same coordinate axis as this bifurcation diagram. You can see the alignment of the main bulbs with the period doubling that occurs along the real axis. The bulb in the "tail" of the Mandelbrot set corresponds to the three-cycle that emerged out of chaos around the value $c = -1.755$ studied earlier.

The Julia sets associated with the c values belonging to the Mandelbrot set vary as the period of the attracting cycle varies among the bulbs. Julia sets for values of c in some of the different bulbs of the Mandelbrot set are shown in Figure 4.40. Notice that the number of "bulbs" in the Julia sets that are pinched together at a single point correspond to the period of the cycles of the Mandelbrot set. For example, the Julia sets for values of c in the main cardioid are all simple closed curves which correspond to the attractive fixed points, whereas the values of c in bulbs that correspond to attractive n-cycles have n bulbs converging at a single point. Notice also the very thin Julia set (dendrite) associated with one of the tendrils of $\mathcal{M}$. Dendrites occur for values of c for which the origin is a periodic point of $f(z) = z^2 + c$; for

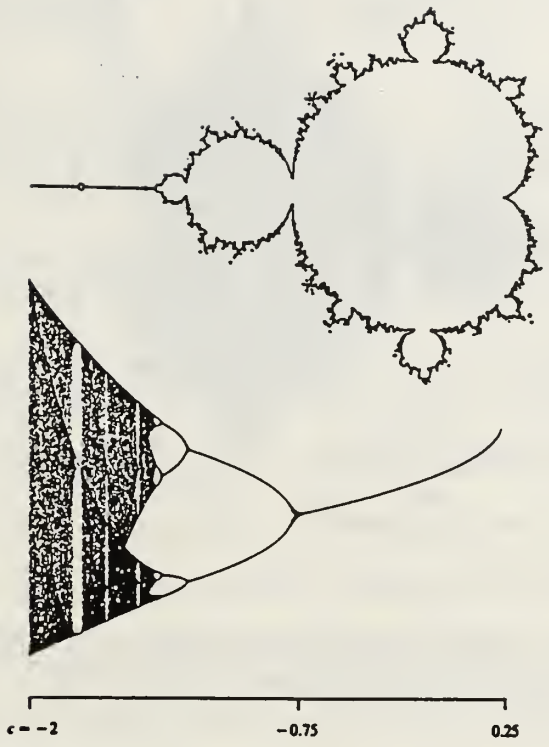

Figure 4.39 The Mandelbrot set plotted against the bifurcation diagram for $f(x) = x^2 + c$. The figure is from the article by Devaney (1989, p. 37).

example the point $c = -i$ as shown in Figure 4.37. Finally, the Julia set associated with a point not in the Mandelbrot set is totally disconnected. Remembering that for values of $c < -2$, the sets of periodic points for the iterated maps $f(x) = x^2 + c$ were Cantor sets (hence totally disconnected) their Julia sets, by definition, are also disconnected. Thus these points fail to belong to the Mandelbrot set.

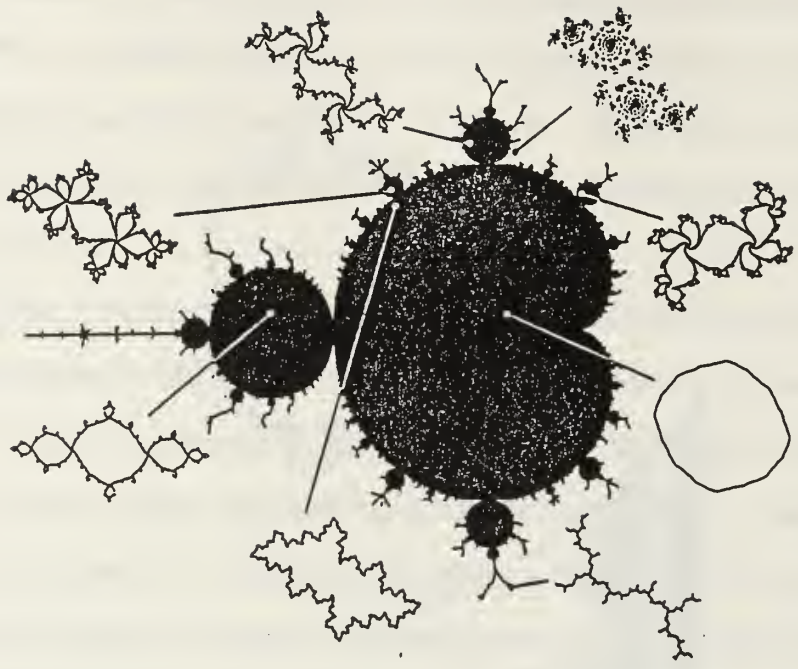

Figure 4.40 The Julia sets of points in different bulbs of the Mandelbrot set (see Figure 4.37). The figure is from Falconer (1990, p. 214).

Another feature of the Mandelbrot set is the existence of a dense set on its boundary of points, called **Misiurewicz points**, for which the image of the Mandelbrot set in parameter space, and the corresponding Julia set in the range space, look the same up to a rotation (in a sense that can be made mathematically precise; see Branner (1989)). Figure 4.41 shows a blowup of the Mandelbrot set and the Julia set around the Misiurewicz point $c \approx -.101096 + i(.956287)$.

The Mandelbrot set occurs in spaces other than the parameter space we have presented. In fact, it appears to be an almost universal geometric shape. Recall the coloring of the complex plane through Newton's method for the function $z^4 - 1 = 0$ in Section IVG. If we color the complex plane for different

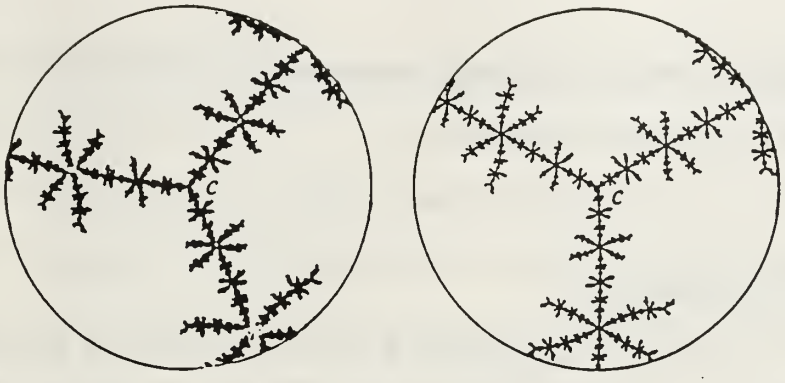

Figure 4.41 The Julia set and the Mandelbrot set around a Misiurewicz point. The figure is from Branner (1989, p. 103).

values of the parameter λ for cubic polynomials of the family P_λ, chaotic regions are found between the basins of attraction. However, interspersed within these regions are small copies of the Mandelbrot set. While these regions have always existed, it has taken present-day powerful computer graphics to reveal them. As scientists continue to use computers to examine dynamical systems more closely, we expect that the Mandelbrot set will appear with increased frequency.

There are, no doubt, other more fascinating properties of the Mandelbrot set yet to be discovered or proven. Each property reveals something about the complexity of the iterated map $f(z) = z^2 + c$. While we have not discussed all known results here, this cursory summary does provide considerable insight into the complexity of this chaotic mapping.

While the function $f(z) = z^2 + c$ appears to be a very specific form of the quadratic family, it is in fact topologically conjugate to every quadratic function for various values of c. To see this, consider the function $H(z) = \alpha z + \beta$ with $\alpha \neq 0$. Then $h^{-1}(f(h(z)) = (\alpha^2 z^2 + 2\alpha\beta z + \beta^2 + c - \beta)/\alpha$. Appropriate choices for the values α, β, and c produce any quadratic

function whatsoever. Thus, in studying the dynamics of $f(z) = z^2 + c$, reveals the entire family of quadratic functions.

N. THE SMALE HORSESHOE

Our attention so far has been restricted to one-dimensional dynamical systems. In so doing, we have learned much about chaos. However, since many real-world phenomena occur in two and three dimensions, the range of applications has been restricted. We now investigate our first two-dimensional dynamical system, the Smale horseshoe. Instead of developing the horseshoe algebraically, a strict geometric interpretation of the map is given. The primary references for this section are Holmes (1989), Guckenheimer (1990), and Devaney (1989).

The Smale horseshoe was originally constructed to help interpret the periodically forced oscillator, which commonly appears in applications in physics, mechanics, and electrical engineering. Normally, the systems under investigation are modeled with ordinary differential equations, and the Smale horseshoe turns out to provide an intuitive way to see why the equations sometimes lead to chaotic behavior.

Many versions of the Smale horseshoe exist. We present here the version that is the simplest geometrically. Thus, take the unit square in Figure 4.42, stretch it out by a factor of three in one direction, and simultaneously shrink it by a factor of three in the other direction to obtain a long bar. Then bend the middle section of the bar into a horseshoe and superimpose it back on the original square, as shown in the figure. Denote this geometric mapping by F. Notice that the two shaded bands do not escape

the unit square under this first iteration. Their preimages are the horizontal bands shown in the figure.

Because the preimage of F can be determined precisely, it is an invertible map. Thus it is possible to study not only the forward orbit of points, but their backward orbits as well. We are interested in finding the invariant set of the unit square under the forward and backward orbits of F. These are the points which do not escape the unit square under infinite forward and backward iteration. Then we will be able to investigate the dynamics of the particular physical system associated with the Smale horseshoe by studying the dynamics on this invariant set.

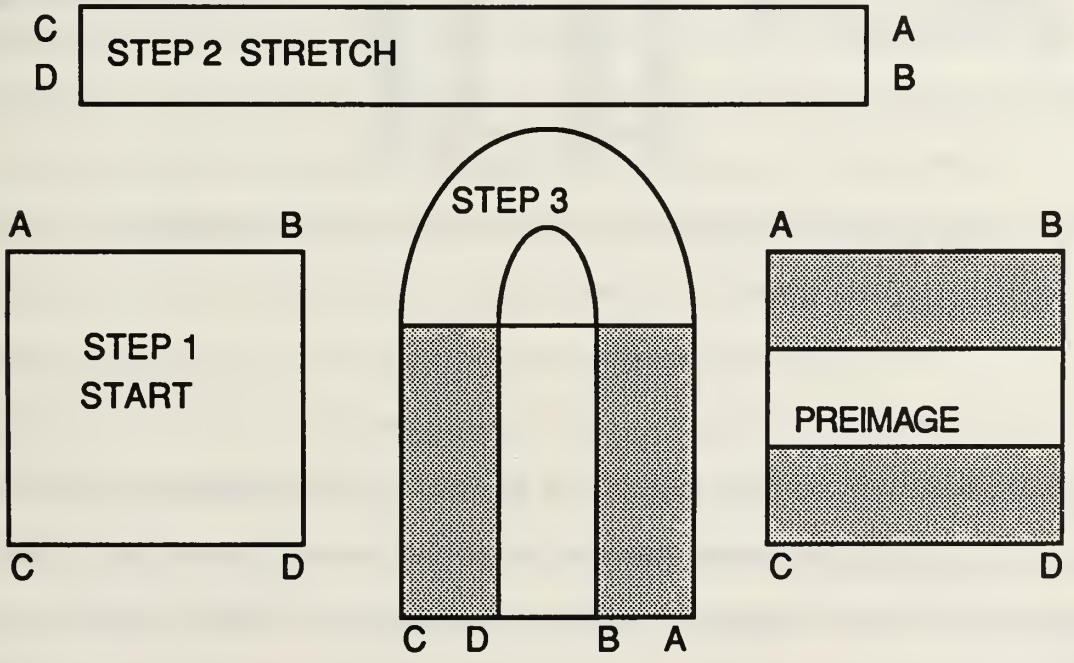

Figure 4.42 Construction of the Smale horseshoe.

Now, iterate the map a second time, as shown in Figure 4.43. Observe that the image of the shaded area of the first iteration, and its preimage appear as before.

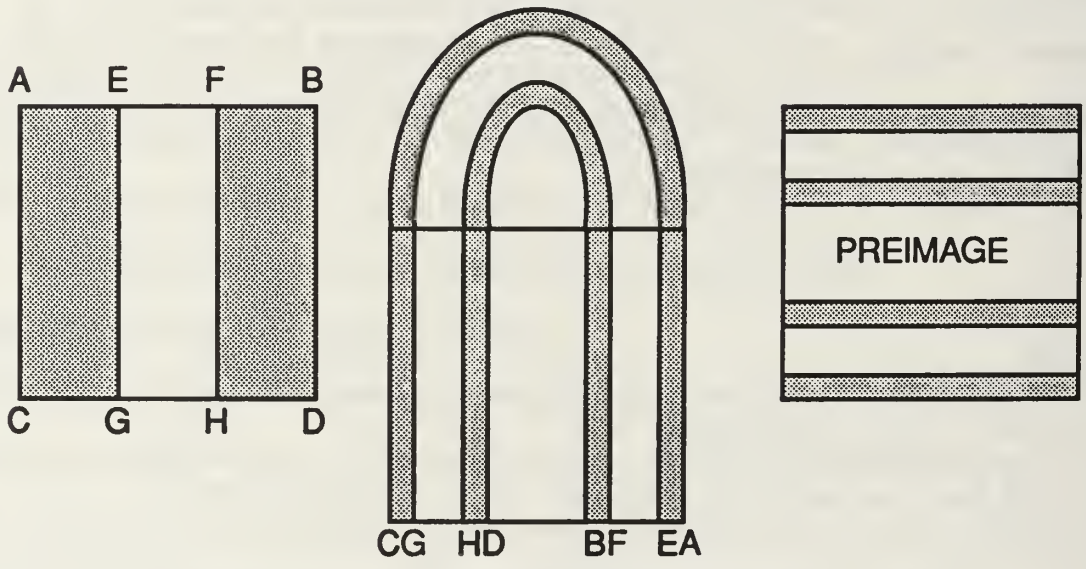

Figure 4.43 The second iteration of the Smale horseshoe.

By superimposing the image of F on its preimage for the first two iterates, we construct geometrically an invariant set, shown as the darkly shaded region in Figure 4.44. Here we label the horizontal and vertical bands H and V, respectively.

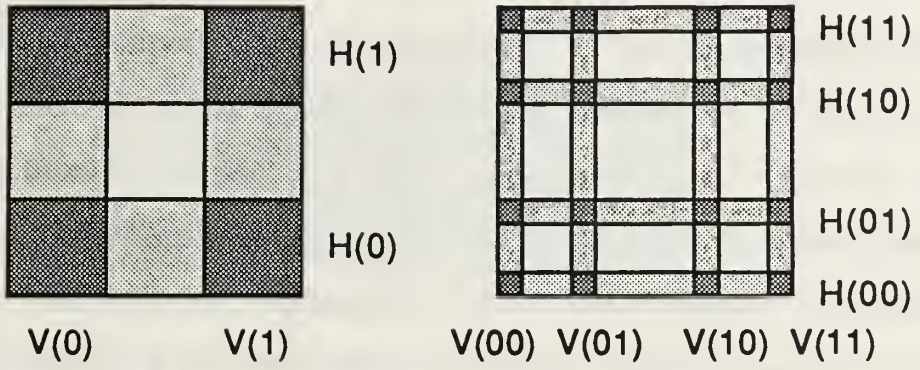

Figure 4.44 The invariant set of the Smale horseshoe.

Notice that the set of points Λ remaining in the unit square under infinite forward and backward iteration has a Cantor-like appearance. In fact, that set turns out to be the direct product of two Cantor middle-thirds sets. The variations of the Smale horseshoe mentioned earlier involve using different values for shrinking and stretching the unit square under F, as well as using a different placement of the horseshoe when it is superimposed back on the square. All variations, however, still create Cantor-like invariant sets.

In order to understand the dynamics of this system, we only need to analyze the dynamics on the invariant set (since all other points escape under iteration for F). To undergo this analysis, first note that the forward and backward orbits of any point x in the invariant set also belong to it. Specifically, each point in these orbits is in one of the horizontal bands H_0 or H_1. Hence, define the mapping $S: \Lambda \rightarrow \Sigma$ by the rule $S_j(x) = i$ if $F_j(x) \in H_i$ for $i \in \{0, 1\}$. Thus, every point x in the invariant set is associated with an infinite string of indices of the horizontal bands to which it is mapped under F. Notice that the index of points in Σ runs $j = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$. So unlike code space Σ_2 for one-dimensional maps (which consisted of semi-

infinite sequences), the space Σ consists of bi-infinite sequences. Figure 4.45 shows a point x and its orbit under three forward iterates and one backward iterate. So for $j = \ldots, -1, 0, 1, 2, 3, \ldots$, $Sj(x) = \ldots 00100 \ldots$ from the horizontal bands in which each iterate lies. Noticing that $S_j(F(x)) = S_{j+1}(x)$, one sees that F applied to the set Λ corresponds to the shift map σ on the space Σ. Moreover, every symbol in Σ corresponds to a unique orbit of F, because every image V completely intersects its preimage H. Therefore, the mapping F and the shift map σ on infinite code space are topologically conjugate through the map S, as shown in Figure 4.46.

The horseshoe map has been very useful in analyzing physical systems because it extends to any Euclidean space R^n. The connection with ordinary differential equations is through a concept known as the **Poincare map**. If the phase space associated with an ordinary differential equation is intersected with a plane normal to any orbit, then the orbit intersects the plane exactly once during each cycle. The collection of these points of intersection is called the **Poincare map**. While the horseshoe map was constructed originally in connection with the Poincare map of a periodically forced oscillator, there is a general method for finding horseshoes that applies to a wide range of Poincare maps. The procedure has helped scientists and engineers understand the dynamics of the associated physical systems.

The actions of stretching and bending in the Smale horseshoe are frequently encountered in physical systems. Predicting the orbit of points in such systems (a simple taffy pull serves as a classical example) has always proven elusive. The science of chaos has helped explain why these systems have been so difficult to understand.

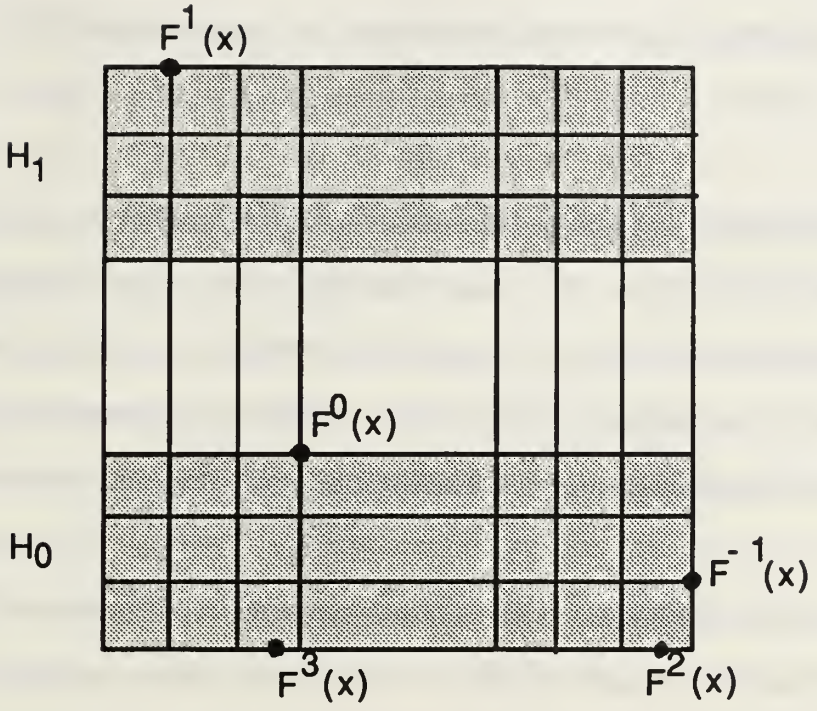

Figure 4.45 The orbit of a point x of the invariant set.

$$\begin{array}{ccc} \Lambda & \xrightarrow{F} & \Lambda \\ \downarrow s & & \downarrow s \\ \Sigma & \xrightarrow{\sigma} & \Sigma \end{array}$$

Figure 4.46 The topological conjugacy between $F: \Lambda \to \Lambda$ and $\sigma: \Sigma \to \Sigma$.

O. THE HENON MAP

With the Smale horseshoe providing a geometric introduction to two-dimensional dynamical systems, we now turn our attention to another map of the plane that exhibits many of the interesting properties of two-dimensional maps. The material in this section is presented as a series of exercises in Devaney (1989), to which most of the answers and results come from Rasband (1990), Alligood (1989), Cherbit (1991) and Moon (1987).

The Henon map $H_{a,b}: \mathbf{R}^2 \to \mathbf{R}^2$, is defined by the equations

$$x_1 = 1 + y_0 - ax_0^2,$$

$$y_1 = bx_0.$$

Notice that H depends on two parameters, a and b, and that it has only one nonlinear term (x^2). Thus H is one of the simplest higher-dimensional nonlinear maps we can study. A number of questions regarding the Henon map have not been resolved because of the wide range of possible parameter values, but for certain parameter values it exhibits some very interesting behavior.

First, note that the Henon map can be expressed as the composition of three maps $H_3 \circ H_2 \circ H_1$, where $H_1(x, y) = (x, 1 - ax^2 + y)$ is a nonlinear bending (and a quick check with calculus shows it is area preserving); $H_2(x, y) = (bx, y)$ is an expansion or contraction in the x direction, depending on the value of b; and $H_3(x, y) = (y, x)$ flips the contracted, bent image about the main diagonal.

The case $b = 0$ makes the Henon map topologically conjugate to the map $g(x) = 1 - ax^2$ if we consider the projection of H onto the x-axis. For the case $|b| > 1$, the map H_2 is not a contraction and the iterates diverge. Hence, we restrict our attention to the range $0 < |b| < 1$.

Now fix b. It is easy to show the fixed points of the Henon map are

$$(x, y) = ([b - 1 \pm ([b - 1]^2 + 4a)^{1/2}]/2a, bx).$$

A closer inspection reveals that for $(b - 1)^2 + 4a < 0$, or $a < -(b - 1)^2/4$, these points have an imaginary component yielding no fixed points in $\mathbf{R}^2$. Moreover, when $a = -(b - 1)^2/4$, the fixed points coincide (so there is only one attracting fixed point). Finally, for $a > -(b - 1)^2/4$, there are two distinct fixed

points, one of which is attracting. Here, for a fixed value of b, as the parameter a increases through a critical parameter value, a tangent bifurcation occurs.

As the parameter a continues to increase a series of period-doubling bifurcations appears eventually leading to chaos. Let a_∞ denote the value of a beyond which chaos occurs. Then the dynamics of the Henon map can be determined geometrically in a familiar setting. For a fixed value of b, let R be the larger root of $a\xi^2 - (b - 1)\xi - 1 = 0$. Let S be the square centered at the origin with vertices at $(\pm R, \pm R)$. Figure 4.47 shows the images of S under H for $a < a_\infty$ and $a > a_\infty$. Note also the effects of H_1, H_2, and H_3 in the way the square S is bent, contracted, and flipped. Additionally, for $a > a_\infty$, the geometric construction looks similar to the Smale horseshoe (and, in fact, it is a horseshoe). Thus the dynamics of the map H for $a > a_\infty$ (for a fixed b) are indeed chaotic.

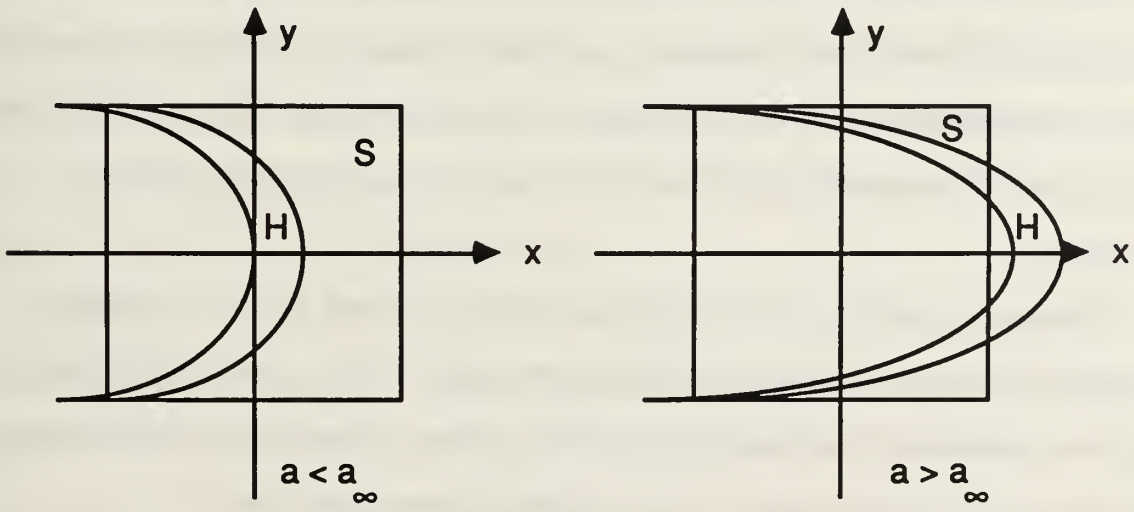

Figure 4.47 The images of S under the Henon map.

The particular value $b = 0.3$ has been studied extensively and the tangent bifurcation occurs at $a = .1225$. If a increases holding $b = 0.3$ constant, in the range $1.052 \leq a \leq 1.082$ a series of period-doubling bifurcations occurs which eventually lead to chaotic behavior.

A particularly interesting phenomenon occurs close to $b = 0.3$ and $a = 1.4$. Here we have an attractor of the system. The infinite iteration of bending, shrinking, and flipping the plane yields results not yet fully understood. Nevertheless, with the aid of computers, it has been possible to compute these results numerically and view them graphically. Iterating an initial point (x_0, y_0) under H yields a set of points, called the **attractor** of H, that appear to be invariant under infinite iteration of H. The attractor of H for the values $b = .3$ and $a = 1.4$ is shown in Figure 4.48. The dynamics on this attractor are chaotic (as just shown geometrically with the analog to the Smale horseshoe). Numerically it has been found that the attractor appears to have a dense orbit, sensitive dependence on initial conditions, and to be topologically transitive. However, since the evidence of this invariant set has only been suggested by the use of numerical computation (and not established with any mathematical rigor) many of its properties are still not clearly identified.

The Henon attractor (if it truly exists) for $b = .3$ and $a = 1.4$ fits into a class of attractors referred to as **strange attractors**. While a formal definition of strange attractor has notbeen developed to date, there are three conditions that seem to be characteristic of them. These characteristics are:

1. Points "nearby" the attractor converge to the attractor under infinite iteration of the function.

2. The dynamics of points on the attractor are chaotic.
3. The attractor has a non-integer fractal dimension.

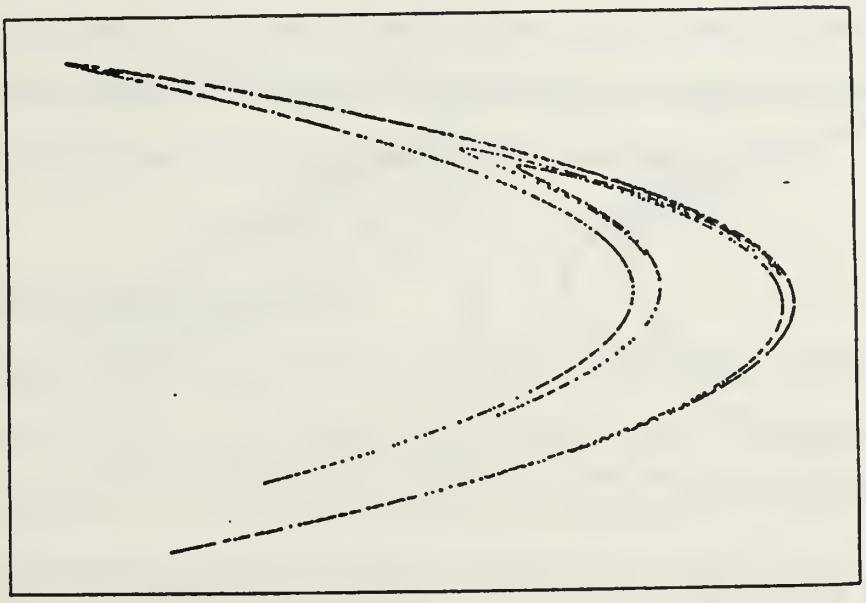

Figure 4.48 The Henon attractor. The figure is from Holden (1986, p. 90).

By "nearby" we refer to a region of the plane called the **basin of attraction,** inside of which all points converge to the attractor. The basin of attraction depends on the particular function, but in the case of the Henon attractor it turns out to be the entire Euclidean plane.

Magnification of the Henon attractor indicates that it is infinitely detailed, as evidenced by the "bands" in Figure 4.49 actually being composed of smaller bands of points. Additionally, its fractal dimension has been estimated numerically at 1.26 for the parameter values b = .3 and a = 1.4. Nevertheless, considerable mystery remains concerning the Henon attractor (as well as many of the other interesting strange attractors that have been

discovered numerically or physically). Because of their structure and self-similarity, fractal geometry is currently being applied to the study of strange attractors.

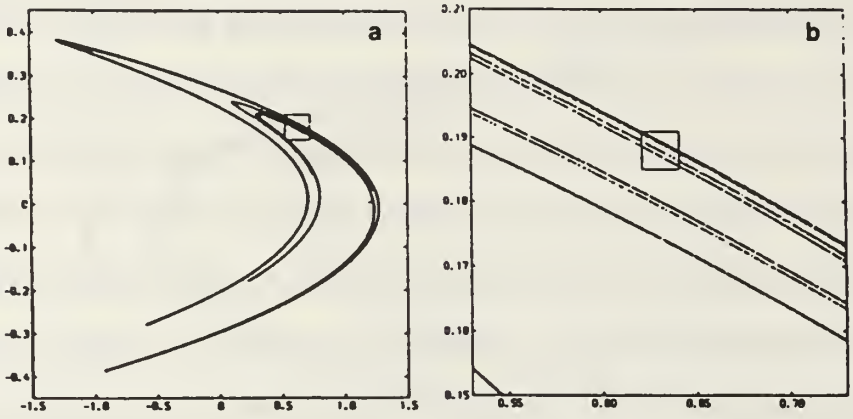

Figure 4.49 Magnification of the Henon attractor. The figure is from Berge (1984, p. 133).

While strange attractors come up in models of physical equations such as the Duffing equation, the van der Pol equations, or the Rossler equations, they have also been seen in physical systems. While many infectious diseases appear to follow definite cycles, measles appears to follow a strange attractor with fractal dimension 2.5 when viewed in the proper phase space. Additionally, Saturn's rings, because of their remarkable resemblance to the strange attractors of many mathematical systems, are being studied in this new light in great detail (however, this connection is still being investigated, and no conclusions have yet been drawn).

A final remark about the Henon map: if we set the parameter $b = 1$, the map becomes an area preserving map of the plane. Since the map $H_1(x,y) =$

$(x, 1 - ax^2 + y)$ is area preserving as observed earlier, we see that with $b = 1$, $H_2(x, y) = (x, y)$ and $H_3(x, y) = (y, x)$ also preserve areas. This condition leads to an entirely new set of phenomena, one of which we briefly mention here. As the parameter a increases, orbits of different periods are created, but the last orbit to develop is a two-cycle. This provides an example of where Sarkovskii's theorem fails to apply in two dimensions.

P. THE LORENZ EQUATIONS

It is appropriate to conclude our mathematical treatment of chaos with the Lorenz equations because they comprise one of the first systems to bring chaotic dynamical systems to the attention of the mathematical community. The primary references for this section are Sparrow (1982), Holden (1986), and Fischer (1985), although some of the presentation follows that of Berge (1984), Guckenheimer (1990), and Thompson (1989).

The Lorenz equations have been studied extensively since the mid 1970s, and numerous interesting results have been derived from them. However, to discuss many of these results requires mathematics beyond the level of this thesis. We present here a cursory summary of some of the results which are consistent with the mathematical level of this thesis, and particularly those which relate to some of the material we have already discussed. A rigorous mathematical derivation of the results we present here can be found in Sparrow (1982).

The Lorenz equations were developed in an attempt to model the earth's atmosphere to simulate weather patterns using a small computer. The Lorenz system is defined as follows:

$$dx/dt = -\sigma x + \sigma y$$

$$dy/dt = rx - y - xz$$

$$dz/dt = xy - bz,$$

with σ, r, and b positive parameters. This is an example of a continuous dynamical system. This system models a flat fluid layer being heated from below and cooled from above (representing the Earth's atmosphere being heated from the ground's absorption of sunlight and losing heat into space). In the resultant temperature flow, x represents the convective motion, y represents the horizontal temperature variation, and z represents the vertical temperature variation. The parameters σ, r, and b are related to the Prandtl number, the Rayleigh number, and the size of the region being modeled (see Figure 4.50).

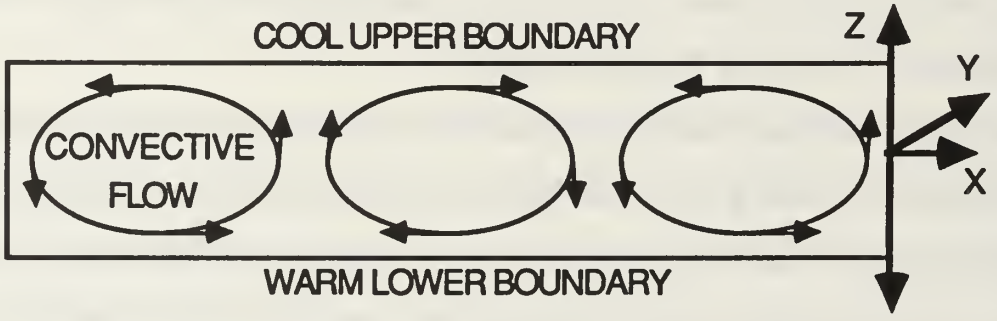

Figure 4.49 The model for the Lorenz equations.

The Lorenz system is a very crude model of weather dynamics and is of little practical value. Actually it has been studied most extensively for parameter values that are nowhere near those of the Earth's atmosphere. While the system does have physical relevance to the Maxwell-Bloch equations for lasers, and to convection problems in specially shaped regions

(usually toroidal), they attract the most attention because of the wealth of information they provide about dynamical systems.

For one-dimensional or planar systems of differential equations, the Poincare-Bendixson theorem (see Hirsch, 1974, p. 248) guarantees that one can completely classify the solution, as to whether it approaches a fixed point or limit cycle, or goes to infinity in a finite amount of time. However, there is no analogous theorem in three dimensions where many systems with interesting behavior are being discovered. The Lorenz system is of great mathematical interest because it possesses many of the characteristics of other higher-dimensional systems. This is not to say it is typical, as it has some distinct characteristics (for example, symmetry) but it does demonstrate characteristics typical of many general higher-dimensional systems.

Because the original paper on this subject by Lorenz (1963) fixed the parameter values at $\sigma = 10$ and $b = 8/3$ and investigated the system as the parameter r varied, much of the literature has taken this same approach, as we do here. Hence, we consider the system

$$dx/dt = 10(y - x),$$

$$dy/dt = rx - y - xz,$$

$$dz/dt = xy - (8/3)z.$$

First note the apparent simplicity of the system. There are only two nonlinear terms, xz and xy. Also, there is a natural symmetry to the equations given by $(x, y, z,) \rightarrow (-x, -y, z)$. The z-axis is invariant because points which start on it stay there and tend towards the origin. Moreover, when $x = 0$ the dx/dt term carries the same sign as y so that all points

which rotate about the z-axis do so in a clockwise manner when viewed from above $z = 0$.

The model has no solutions which tend to infinity; in other words, there is a surface inside of which all solutions tend towards the origin and herein all solutions remain. To see this, consider the ellipsoid $f(x, y, z) = x^2/2\sigma + y^2/2 + z^2/2 - (r + 1)z - \mu = 0$ for μ arbitrarily large. We show that the dot product of the velocity vector and the outward normal vector to the ellipsoid is always negative, i.e., $Df = (dx/dt)f_x + (dy/dt)f_y + (dz/dt)f_z < 0$. Substituting into the Lorenz equations we obtain $Df = \sigma(y - x)f_x + (rx - y - xz)f_y + (xy - bz)f_z$. Then substitution of the ellipsoid partial derivatives yields $Df = -x^2 - y^2 - bz^2 + (r + 1)bz$. For large enough μ in the equation for the ellipsoid, the quadratic term in z in the expression for Df always dominates the linear term in z, so for this surface the flow is always towards the origin. Hence, no trajectory originating a finite distance from the origin will go to infinity.

We now analyze the system for $\sigma = 10$ and $b = 8/3$ as we vary r. To begin, we restrict our attention to small values of r (i.e., $r < 30$). A quick check shows the origin is a fixed point for all parameter values, but we would like to know whether it is attracting or repelling.

We introduce here the concepts of *stable* and *unstable manifolds*. A **stable manifold** of a point p is the set of all points that tend to p in forward time (as $t \to \infty$). The **unstable manifold** of p is the set of points that tend to p in backward time (as $t \to -\infty$). For our purposes, a **manifold** can be thought of as simply a surface in phase space. These manifolds can be determined from the eigenvalues of the linearized system near the point p. The linearized system has the matrix:

$$\begin{bmatrix} -\sigma & \sigma & 0 \\ r-z & -1 & -x \\ y & x & -b \end{bmatrix},$$

which, when evaluated at the origin yields

$$\begin{bmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{bmatrix}.$$

The eigenvalues of this system are $\lambda_1, \lambda_2 = (1/2)[-\sigma - 1 \pm ((\sigma - 1)^2 + 4\sigma r)^{1/2}]$, and $\lambda_3 = -b$.

Since, for $r < 1$, we have $[(\sigma - 1)^2 + 4\sigma r]^{1/2} < [(\sigma - 1)^2 + 4\sigma]^{1/2} = \sigma + 1$, so all three eigenvalues of the linearized system evaluated at the origin are negative. Hence, the origin is globally attracting. The phase portrait of this condition is shown in Figure 4.51. However, for $r = 1$, the eigenvalues evaluated at the origin are $\lambda_1 = 0$, $\lambda_2 = -\sigma - 1$, and $\lambda_3 = -b$. This zero eigenvalue is analogous to the nonhyperbolic fixed point we encountered in our study of discrete systems. We require more theory to determine whether the manifold associated with this eigenvalue is stable or unstable. For $r > 1$, the eigenvalues are $\lambda_1 > 0$, $\lambda_2 < 0$, and $\lambda_3 < 0$. Since two of these eigenvalues are negative and one is positive, the origin has a two-dimensional stable manifold and a one-dimensional unstable manifold for $r > 1$.

As r passes through 1, not only does the origin become unstable, but two new fixed points are introduced at $(\pm b(r - 1)^{1/2}, \pm b(r - 1)^{1/2}, r - 1)$ which we denote C^+ and C^-. You should recognize this as a bifurcation. A similar check as we did above of the linearized system near these points shows that they

have complex eigenvalues. Furthermore, for values of $r < r_H$ (as defined below), the real parts are all negative. Hence, these points are attracting. The phase portrait of the system for $1 < r < r_H$ is shown in Figure 4.52.

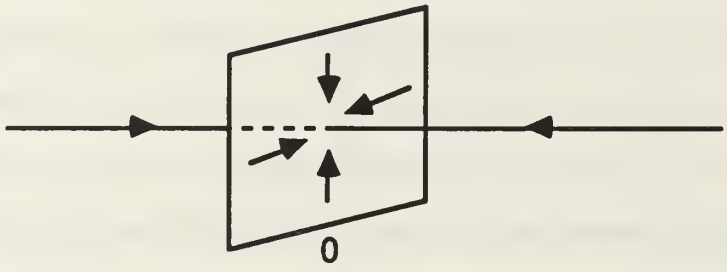

Figure 4.51 The origin is globally attracting for $r < 1$.

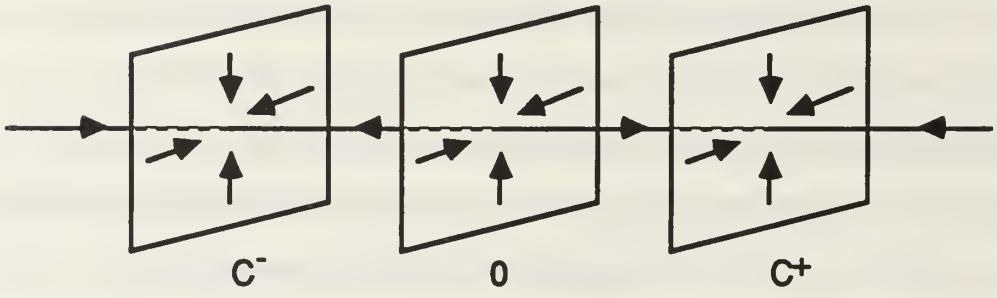

Figure 4.52 Phase portrait of the Lorenz system for $1 < r < r_H$.

Numerical solutions to the Lorenz equations indicate that for $1 < r < 13.926$, orbits on the unstable manifold of the origin tend directly to the nearest attracting fixed point C^+ or C^-, as indicated in Figure 4.53. However, for $r > 13.962$, these orbits "cross over" and are attracted to the other stable point (see Figure 4.54).

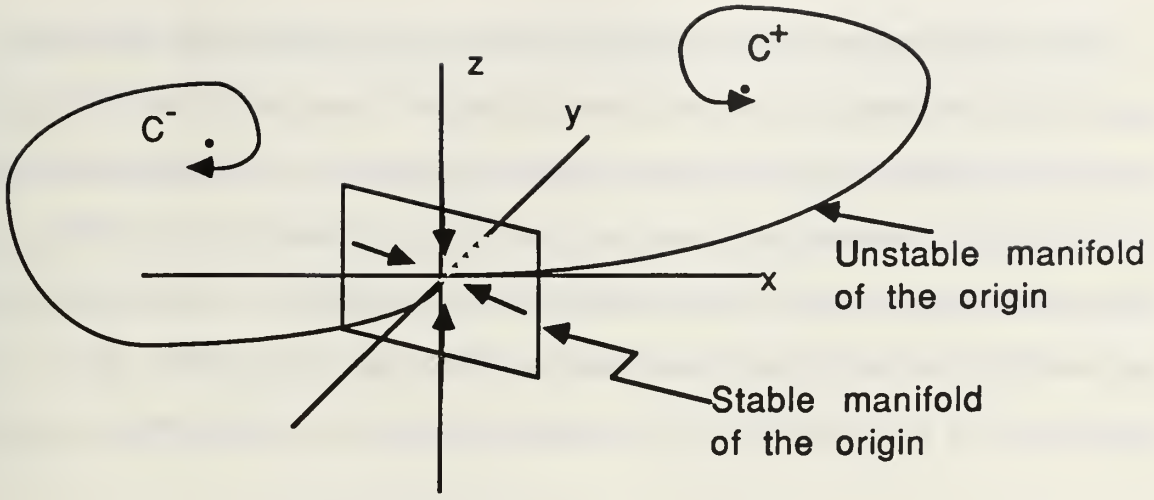

Figure 4.53 Solution trajectories for $r < 13.962$.

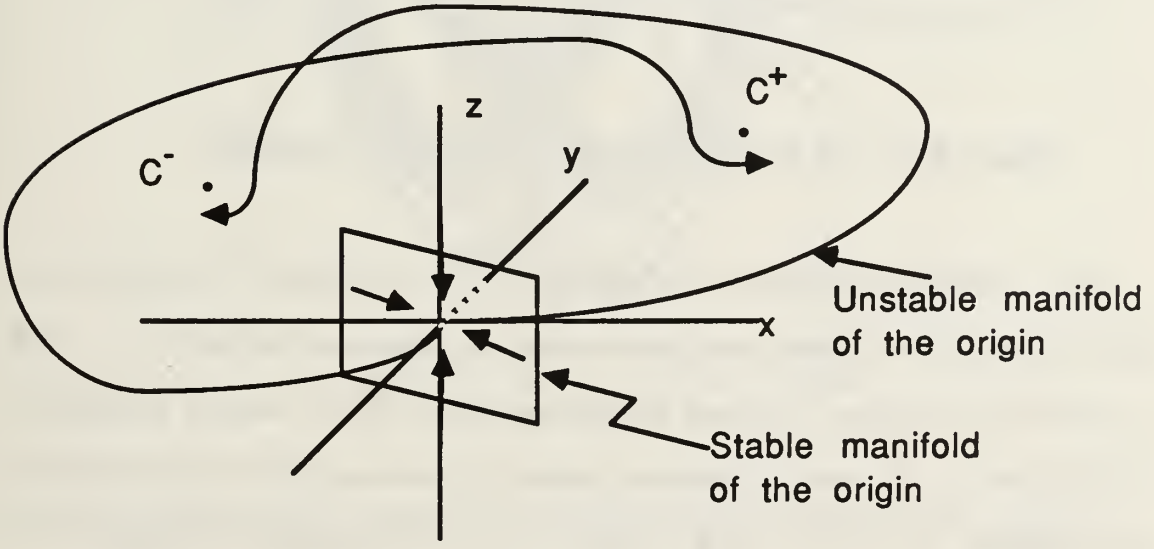

Figure 4.54 Solution trajectories for $r > 13.962$.

Since we know the stable manifold of the origin is planar near the origin and includes the entire z-axis, and since trajectories cannot cross each other, the stable manifold must be twisted in some strange way. What happens here is the stable and unstable manifolds of the origin merge and form an orbit called a homoclinic orbit. A **homoclinic orbit** of a point p is a set of points that tend to p in both forward and backward time (see Figure 4.55). The introduction of a homoclinic orbit is another example of a bifurcation.

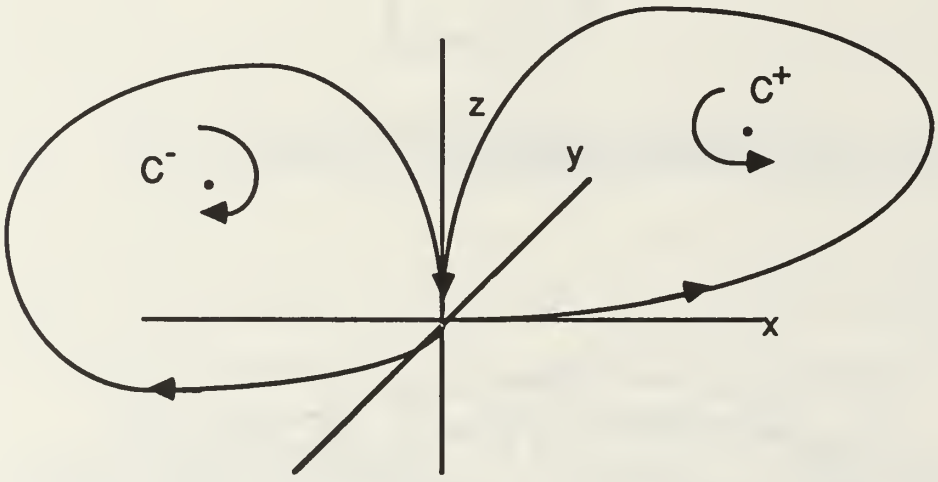

Figure 4.55 The homoclinic orbit of the Lorenz equations.

As r continues to increase, we note that at $r = r_H = [\sigma(\sigma + b + 3)]/(\sigma - b - 1)$, the real parts of the complex eigenvalues of the linearized system at C^+ and C^- cross the imaginary axis and become positive. This is another example of a bifurcation as C^+ and C^- become unstable. Hence all three fixed points are now repelling. For $\sigma = 10$ and $b = 8/3$, $r_H \approx 24.74$, numerical solutions indicate that for $r > r_H$, there is an attractor (called the invariant set) to which all solutions tend as $t \to \infty$.

Figure 4.56 shows this invariant set for $r = 28$, $b = 8/3$, and $\sigma = 10$ as it is projected onto the xz-plane. This invariant set is a strange attractor and it exhibits some interesting properties. For example, the trajectory continues forever within the bounds shown, yet never crosses itself or returns to the same point in space. Additionally, the dynamics on the attractor are believed to be chaotic, although for continuous systems more theory is required than developed in this thesis.

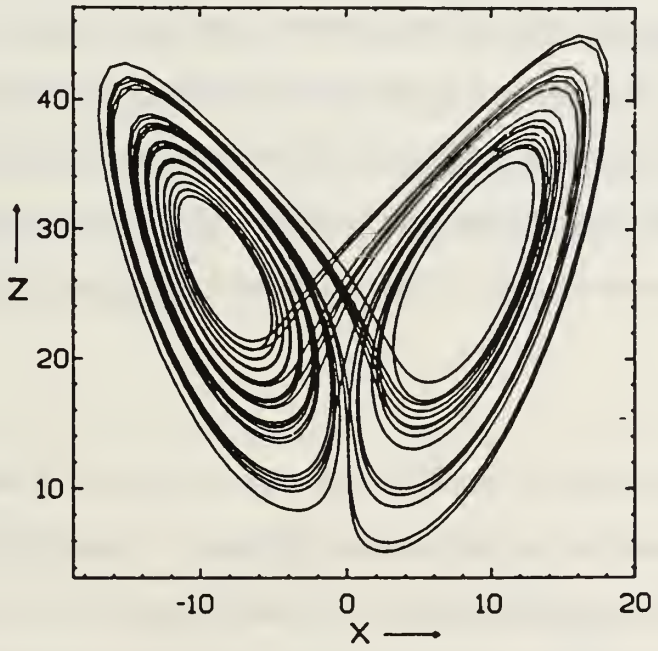

Figure 4.56 The Lorenz attractor. The figure is from Holden (1986, p. 126).

There is no closed form solution to the Lorenz equations, so most of the evidence as to their behavior has been obtained numerically and is

conjectured. However, the following results for small values of r have been verified computationally numerous times, and are widely accepted as the system's true behavior:

1. For $r < 1$, all solutions tend towards the origin.
2. For $1 < r < 13.926$, all trajectories spiral into one of the attracting fixed points C^+ and C^-.
3. For $13.962 < r < 24.06$ an invariant set appears in the trajectory, and some solutions wander among the invariant set before spiraling into either C^+ or C^-. The closer r gets to 24.06, the longer some solutions stay near the invariant set.
4. For $r > 24.06$, some trajectories stay forever near the invariant set, although for $r < 24.74$, some trajectories eventually spiral into C^+ or C^-. For $r > 24.74$, the fixed points C^+ and C^- become repelling, and all trajectories remain forever near the invariant set. These invariant sets are similar to the one shown in Figure 4.56, and get closer to it as r increases.

Using an analysis similar to the Poincare map, we can see a further connection between continuous and discrete systems. Considering the homoclinic orbit of the Lorenz equations, we can construct a small box about the origin, and analyze where orbits near the homoclinic orbit penetrate this surface. This analysis shows variations on many of the exotic structures we studied for discrete systems, including horseshoes and Cantor "books" or "fans," which are families of two-dimensional Cantor sets "sewn" together

along a one-dimensional manifold, or "spine." These intriguing results are all presented in Sparrow (1982).

For large values of r, numerical solutions to the Lorenz equations have been obtained that exhibit many of the phenomena we studied earlier for discrete systems. For values of r in certain intervals (called "windows"), stable periodic orbits develop that bifurcate as r decreases through the window. One such window appears at $99.542 < r < 100.795$. For $99.98 < r < 100.795$, the orbit shown in Figure 4.57 appears. In the interval $99.62 < r < 99.98$, a different orbit appears which has two "loops" that pass very close to each other (see Figure 4.57). This is a period-doubling bifurcation as r decreases through the critical value $r = 99.98$. An entire sequence of period-doubling bifurcations occur as r decreases from 100.795 to 99.542. If we let r_n be the values at which these bifurcations occur, then evaluation of the ratio $(r_{n-1} - r_n)/(r_n - r_{n+1})$ yields approximately 4.67 in the limit, which is very close to the Feigenbaum constant.

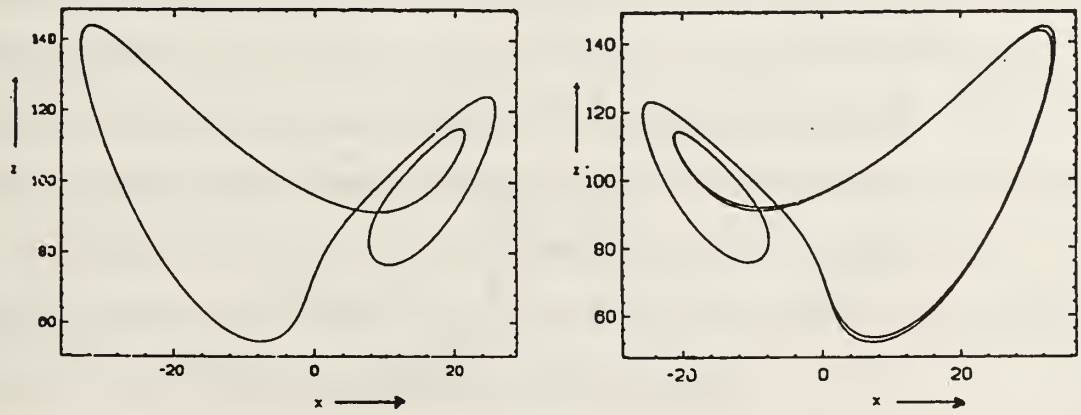

Figure 4.57 Orbits for $r = 100.5$ and $r = 99.65$. The figure is from Sparrow (1982).

A second window appears from $145 < r < 166$. The numerical solutions for $r = 160$ and $r = 147.5$ are shown in Figure 4.58. Again, period-doubling is apparent. A final window occurs for $214.364 < r < \infty$. Period-doubling is again observed in the solutions for $r = 350, 260, 222$, and 216.2, as shown in Figure 4.59. Additionally, a symmetric orbit is seen for $r = 350$.

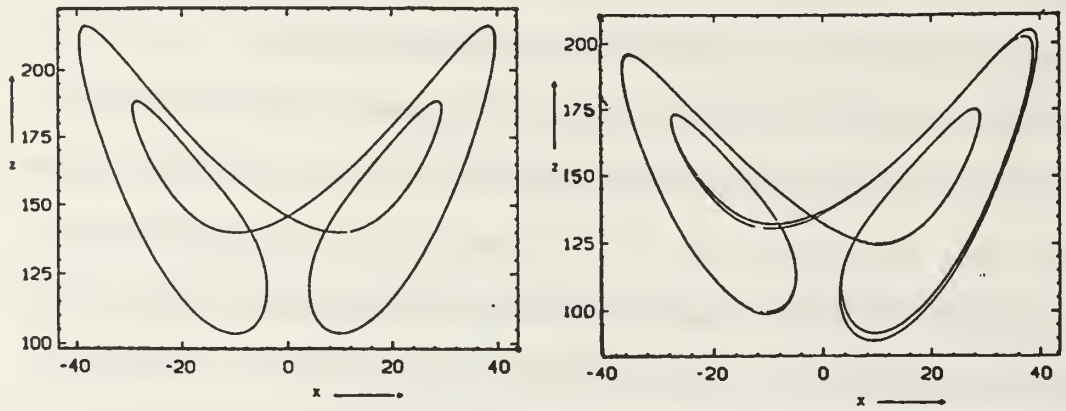

Figure 4.58 Orbits for $r = 160$ and $r = 147.5$, showing period-doubling. The figure is from Sparrow (1982).

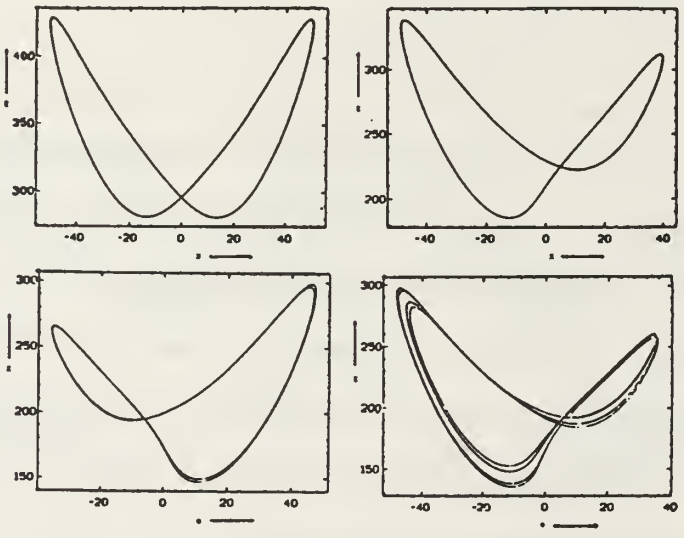

Figure 4.59 Orbits for $r = 350, 260, 222$, and 216. The figures are from Sparrow (1982).

These are just a few examples of the many observed phenomena of the Lorenz equations. Additionally, because of the wide range of parameter values, there are even more unanswered questions about the system. However, in light of what we studied for discrete dynamical systems, these results have a direct analogy to the discrete phenomena we studied earlier.

Although the Lorenz equations reveal very little about the weather, they do give considerable information concerning continuous dynamical systems, a small amount of which was discussed here.

The behavior of the system does tell us that weather is unpredictable to any degree of accuracy projected for any large amount of time into the future. From what we know about chaotic dynamical systems (and weather is surely chaotic), even if we were to develop an accurate model and measure atmospheric conditions accurately on an arbitrarily small grid, the sensitive dependence on initial conditions of the system causes any computed (predicted) solution to stray arbitrarily from the actual weather, given the slightest reading error. Additionally, small perturbations (which could never be modeled) such as a single person lighting a match, could cause the whole system to follow a new orbit. On the other hand, if weather follows some strange attractor on which the dynamics of the system are chaotic, then not only is it unpredictable (sensitive dependence on initial conditions), but every type of weather possible (topological transitivity) could be experienced. Moreover, there will be a dense period, providing some order to the weather allowing us to predict such things as seasonal changes.

LIST OF REFERENCES

Alligood, Kathleen T., and James A Yorke, "Fractal Basin Boundaries and Chaotic Attractors," *Proceedings of Symposia in Applied Mathematics,* Volume 39, American Mathematical Society, 1989.

Anton, Howard, and Chris Rorres, *Elementary Linear Algebra with Applications,* John Wiley & Sons, 1987.

Arnold, V. I., *Ordinary Differential Equations,* The Massachusetts Institute of Technology Press, 1989.

Barnsley, Michael, *Fractals Everywhere,* Academic Press, Inc., 1988.

Berge, Pierre, Yves Pomeau, and Christian Vidal, *Order Within Chaos,* John Wiley & Sons, 1984.

Boas, Ralph P. *Invitation to Complex Analysis,* Random House Inc., 1987.

Branner, Bodil, "The Mandelbrot Set," *Proceedings of Symposia in Applied Mathematics,* Volume 39, American Mathematical Society, 1989.

Briggs, John and F. David Peat, *Turbulent Mirror,* Harper & Row, 1989.

Cherbit, Guy, *Fractals,* John Wiley & Sons, 1991.

Churchill, Ruel V. and James W. Brown, *Complex Variables and Applications,* McGraw-Hill Publishing Co., 1990.

Devaney, Robert L., *Chaotic Dynamical Systems,* Addison-Wesley Publishing Co., Inc., 1989.

Devaney, Robert L., "Dynamics of Simple Maps," *Proceedings of Symposia in Applied Mathematics,* Volume 39, American Mathematical Society, 1989. (Cited as "the article by Devaney.")

Devaney, Robert L., "The Orbit Diagram and the Mandelbrot Set", *The College Mathematical Journal,* Vol 22, No 1, Mathematical Association of America, Jan, 1991.

Falconer, Kenneth, *Fractal Geometry,* John Wiley & Sons, 1990.

Fischer, Ismor, "The Lorenz Equations: A Short Analysis," Paper submitted in conjunction with a course given at the University of Wisconsin, Spring, 1985.

Giordano, Frank R., and Maurice D. Weir, *Differential Equations, A Modeling Approach*, Addison-Wesley Publishing Co., 1991.

Gleick, James, *Chaos, Making a New Science*, Penguin Books, 1988.

Guckenheimer, John, and Philip Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, 1990.

Harrison, Jenny, "An Introduction to Fractals," *Proceedings of Symposia in Applied Mathematics*, Volume 39, American Mathematical Society, 1989.

Hirsch, Morris W. and Stephen Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, 1974.

Holden, Arun V., *Chaos*, Princeton University Press, 1986.

Holmes, Philip, "Nonlinear Oscillations and the Smale Horseshoe Map," *Proceedings of Symposia in Applied Mathematics*, Volume 39, American Mathematical Society, 1989.

Keen, Linda, "Julia Sets," *Proceedings of Symposia in Applied Mathematics*, Volume 39, American Mathematical Society, 1989.

Levy, Steven, "It's Alive," *Rolling Stone*, Issue 606, June 13, 1991.

Lorenz, E. N., "Deterministic Non-Periodic Flows," *Journal of Atmospheric Science*, Volume 20, 1963.

Moon, Francis C., *Chaotic Vibrations*, John Wiley & Sons, 1987.

Rasband, S. Niel, *Chaotic Dynamics of Nonlinear Systems*, John Wiley & Sons, 1990.

Ross, Kenneth A., *Elementary Analysis: The Theory of Calculus*, Springer-Verlag, 1980.

Seydel, Rudiger, *From Equilibrium to Chaos*, Elsewier, 1988.

Sparrow, Colin, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, Springer-Verlag, 1982.

Strang, Gilbert, "A Chaotic Search for i," *The College Mathematical Journal*, Vol 22, No 1, Mathematical Association of America, Jan, 1991.

Thompson, J. M. T., and H. B. Stewart, *Nonlinear Dynamics and Chaos*, John Wiley & Sons, 1989.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, VA 22304-6145	2
2.	Library, Code 52 Naval Postgraduate School Monterey, CA 93943-5002	2
3.	Department Chairman, Code MA/Fs Department of Mathematics Naval Postgraduate School Monterey, CA 93943-5000	1
4.	Professor M. Weir, Code MA/Wc Department of Mathematics Naval Postgraduate School Monterey, CA 93943-5000	1
5.	Professor I. Fischer, Code MA/Fi Department of Mathematics Naval Postgraduate School Monterey, CA 93943-5000	1
6.	Li, Hsin-Yun 2F, No. 9, Lane 21, Minli S.T. Gushan Kaohsiung Taiwan, Republic of China	1
7.	Beaver, Philip P.O. Box 1230 Santa Barbara, CA, 93102	2

8. Van Joolen, Vincentius — 1
LT, USNR
Department Head Class 120
Surface Warfare Officers School Command
Newport, RI 02841-5012

9. Professor W. Colson, Code PH/Cl — 1
Department of Physics
Naval Postgraduate School
Monterey, CA 93943-5000

10. Professor A. Schusteff, Code MA/Sh — 1
Department of Mathematics
Naval Postgraduate School
Monterey, CA 93943-5000

11. Chairman — 1
Department of Mathematics
United States Military Academy
West Point, NY 10996-1786

12. Dean of Faculty and Graduate Studies — 1
Code 07
Naval Postgraduate School
Monterey, CA 93943-5000

Thesis
B3175 Beaver
c.1 Fractals and chaos.